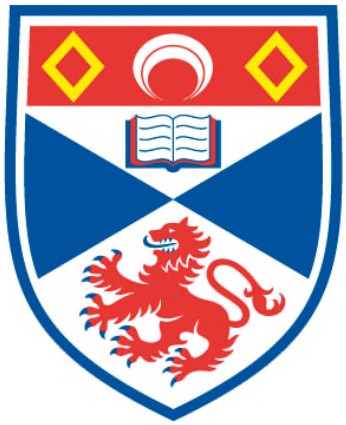
THE ROLE OF CTCF IN THE LIFE CYCLE OF HUMAN PAPILLOMAVIRUS

Christian Paris

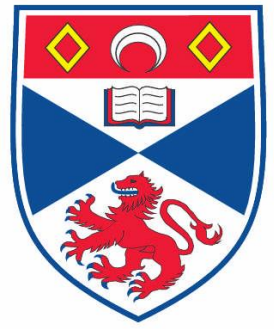
A Thesis Submitted for the Degree of PhD
at the
University of St Andrews



2014

The Role of CTCF in the Life Cycle of Human Papillomavirus

Christian Paris



This thesis is submitted in partial fulfilment for the degree of Doctor of Philosophy at the

University of St Andrews
School of Medicine

May 2014

Abstract

Papillomaviruses (PV) are epithelium specific DNA viruses that can cause health problems ranging from harmless warts to invasive cancer. Papillomavirus induced tumours most often arise in the cervix where human papillomavirus (HPV) infections were shown to cause 99.7 % of all malignancies.

This study aims to map binding sites of the multifunctional host protein CCCTC binding factor (CTCF) to the papillomavirus genome, validate them and determine the function of CTCF in the papillomavirus life cycle. Computer predictions of CTCF binding sites in the sequence of 8 different PV revealed a CTCF binding pattern including a conserved high-affinity binding site around nucleotide 3000 in high risk HPV and around nucleotide 5400 in low risk HPV. This binding pattern was experimentally confirmed using electrophoretic mobility shift assays (EMSA). The binding site around nucleotide 3000 in HPV18 was mutated and human foreskin keratinocytes (HFK) were transfected with mutant and wild type HPV18 to analyse the effect of the mutation on viral gene expression and life cycle. Western blotting of methylcellulose differentiated HFK revealed earlier expression of E2 and decreased expression of E1^E4 in the mutant compared to the wild type. Immunostaining of organotypic raft cultures grown from the mutant maintaining cells showed a significant increase in proliferating cells compared to the HFK maintaining the wild type. This was accompanied by pseudo-differentiation of keratinocytes since the cells of the granular layer of the raft expressed the terminal differentiation marker loricrin but maintained the morphology of undifferentiated cells. Thus CTCF was shown to have a major impact on the HPV life cycle and it may play a role in HPV induced carcinogenesis. Furthermore a function of CTCF in long term maintenance of the viral episome was revealed as cells maintaining the CTCF mutant were shown to lose episomes more quickly compared to wild type maintaining cells.

Declarations

1. Candidate's declarations:

I, Christian Paris, hereby certify that this thesis, which is approximately 65000 words in length, has been written by me, and that it is the record of work carried out by me, or principally by myself in collaboration with others as acknowledged, and that it has not been submitted in any previous application for a higher degree.

I was admitted as a research student in November 2009 and as a candidate for degree of Doctor of Philosophy in November 2010; the higher study for which this is a record was carried out in the University of St Andrews between 2009 and 2014.

Date _____ Signature of candidate _____

2. Supervisors' declaration:

We hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of Doctor of Philosophy in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

Date _____

Signature of supervisor, University of St Andrews _____

Date _____

Signature of supervisor, University of Birmingham _____

3. Permission for publication:

In submitting this thesis to the University of St Andrews I understand that I am giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. I also understand that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that my thesis will be electronically accessible for personal or research use unless exempt by award of an embargo as requested below, and that the library has the right to migrate my thesis into new electronic forms as required to ensure continued access to the thesis. I have obtained any third-party copyright permissions that may be required in order to allow such access and migration, or have requested the appropriate embargo below.

The following is an agreed request by candidate and supervisor regarding the publication of this thesis:

Embargo on both [all] of printed copy and electronic copy for the same fixed period of 3 years on the following ground: publication would preclude future publication.

Date _____ Signature of candidate _____

Date _____

Signature of supervisor, University of St Andrews _____

Date _____

Signature of supervisor, University of Birmingham _____

Acknowledgements

First I would like to thank my supervisors Jo Parish and Simon Powis for numerous different things that can be summarised as great support. I want to thank Jo for her tireless and invaluable effort in discussing and improving my experiments and for the continued support after her move to Birmingham. Simon was taking over my supervision at St Andrews after Jo had moved and his opinions, suggestions and his ability to motivate me had a major impact on the success of this study.

Just like supervision, support and teamwork in the lab is crucial for any research project, including this project. I want to thank Katherine Feeney and Chris Wasson for introducing me to the first steps in the new lab. Also the other members of the Parish lab were always happy to support me and I am very thankful for that. These members are Laura McFarlane-Majeed, Leanne Harris, and the former members of the Parish group, Kris Okrasa and Arif Sheikh. Additionally I wish to thank Elaine Campbell who was always there for me when help or advice was needed.

Also people from other Universities contributed to this thesis and I greatly appreciate their effort. Dawn Farrar supplied me with the plasmids needed for CTCF expression. Jesse Ziebarth was always happy to discuss any question about the predictions of CTCF binding sites and he provided me with the raw data from *Storm*. Ian Groves provided data for HPV16 and supported my ChIP assays. I also greatly appreciate the help of Sally Roberts who provided the expertise for generating HPV18 maintaining human foreskin keratinocytes.

Finally I would like to thank James McDonagh, Jasmin Schnick and Rob Shore for spicing up my time in St Andrews with beer, banter and squash.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Acronym	Definition
ΔN63α	Isoform of epidermal growth factor receptor
°C	Degrees Celsius
µg	Microgram
µl	Microliter
AB	Antibody
AC	Adenocarcinoma
ADP	Adenosine diphosphate
APS	Ammoniumpersulfate
***Bam*HI**	Restriction endonuclease from *Bacillus amyloli*
***Bgl*II**	Restriction endonuclease from *Bacillus subtilis*
BORIS	Brother of the regulator of imprinted sites (identical to CTCFL)
bp	Base pairs
BPV	Bovine papillomavirus
BrdU	Bromodeoxyuridine
BSA	Bovine serum albumin
C3	Mutant of CTCF binding site 2990 in HPV18
CaSki	Cervical cancer cell line transformed by HPV16
CC3	Clonetics human foreskin keratinocytes maintaining C3 mutant HPV18
CDP	CCAAT-displacement protein
ChIA-PET	Chromatin interaction analysis with paired end tag sequencing
ChIP	Chromatin immunoprecipitation
CIN1	Cervical intraepithelial neoplasm grade 1
CIN2	Cervical intraepithelial neoplasm grade 2
CIN3	Cervical intraepithelial neoplasm grade 3
CIN612	HPV31b maintaining cervical intraepithelial neoplasm cell line
CR1	Conserved region 1
CR2	Conserved region 2
Ct	Cycle threshold
CTCF	CCCTC binding factor
CTCFBSDB	CTCF binding site database
CTCFBSDB 1.0	Version 1.0 of the CTCF binding site prediction tool of the CTCF binding site database
CTCFBSDB 2.0	Version 2.0 of the CTCF binding site prediction tool of the CTCF binding site database
CTCFL	CCCTC-binding factor-like (identical to BORIS)
CWT	Clonetics human foreskin keratinocytes maintaining wild type HPV18
ddH_2O	Double distilled Water
DEK	Chromatin associated protein that is abundant in proliferating cells

DNA	Deoxyribonucleic acid
DNMT1	DNA methyltransferase 1
***Dpn*I**	Methylation-specific nuclease
DTT	Dithiothreitol
E1	Papillomavirus early antigen 1
E1^E4	Papillomavirus early antigen 4
E2	Papillomavirus early antigen 2
E2-TR	Papillomavirus early antigen 2 transcriptional repressor
E5	Papillomavirus early antigen 5
E6	Papillomavirus early antigen 6
E6AP	E6-associated protein
E7	Papillomavirus early antigen 7
E8^E2	Repressor variant of BPV1 early antigen 2
E8^E2C	Repressor variant of alpha HPV early antigen 2
EBER1	Epstein-Barr virus encoded RNA 1
EBER2	Epstein-Barr virus encoded RNA 2
EBNA1	Epstein-Barr nuclear antigen 1
EBNA2	Epstein-Barr nuclear antigen 2
EBV	Epstein-Barr virus
***Eco*RI**	Restriction endonuclease from *Escherichia coli*
EDTA	Ethylenediaminetetraacetic acid
ER	Estrogen receptor (alpha)
FAM	Fluorescein amidite
FASTA	Text-based format for representing nucleotide and protein sequences
FLAG	Polypeptide protein tag
fmol	Femtomol
FOXA1	Forkhead box protein A1
FW	Forward
g	Gram
G2 -phase	Second growth phase of the cell cycle
GAPDH	Glyceraldehyde 3-phosphate dehydrogenase
GATA 3	GATA binding protein 3
GC3	Georgie human foreskin keratinocytes maintaining C3 mutant HPV18
GWT	Georgie human foreskin keratinocytes maintaining wild type HPV18
H_2O	Water
H_2O_2	Hydrogen peroxide
HaCaT	Cultured human keratinocytes
HCl	Hydrochloric acid
hDLG	Human drosophila disc large tumor suppressor
HeLa	Helen Lane (Henrietta Lacks), a cell line of HPV18-induced cervical carcinoma
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HFK	Human foreskin keratinocyte

HINGS	Heat-inactivated goat serum
Hi-C	Method for the investigation of 3-dimentional genome architecture
HIS-tag	Hexahistidine-tag
HPV	Human papillomavirus
hScribble	Tumour suppressor protein targeted for degradation by E6
HSIL	High grade squamous intraepithelial lesion
HSV1	Herpes simplex virus type 1
hTERT	Human telomerase
HVS	Herpesvirus Saimiri
ICP0	Promoter of herpes simplex virus type 1
ICP4	Promoter of herpes simplex virus type 1
ICR	Imprinting control region
ID13	Mouse fibroblast cell line maintaining BPV1
Igf2/H19	Model locus for the role of CTCF enhancer blocking, imprinting and methylation dependent gene regulation by CTCF
IMS	Industrial methylated spirits
KCL	Potassium chloride
kDa	Kilo Dalton
KDM6A	Lysine (K)-specific demethylase 6A
KDM6B	Lysine (K)-specific demethylase 6B
KSHV	Kaposi's sarcoma associated herpesvirus
l	Litre
L1	Papillomavirus late antigen 1
L2	Papillomavirus late antigen 2
LAD	Lamina associated domain
LAT	Latency associated transcript
LB broth	Luria-Bertani broth
LCR	Long control region
LSIL	Low grade squamous intraepithelial lesion
M	Molar
M13	DNA sequence, overhang of forward primers for generating DNA fragments for EMSA
MAGI-1	Membrane-associated guanylate kinase, WW and PDZ domain-containing protein 1
mg	Milligram
MgCl2	Magnesium chloride
min	Minute
mM	Millimolar
mm	Millimeters
mRNA	Messenger ribonucleic acid
MUPP1	Multiple PDZ domain protein
MW	Molecular weight

MYC	Proto-oncogene C-Myc
Myb	Myeloblastosis viral oncogene homolog
MyoD	Myogenic Differentiation 1
NaCl	Sodium chloride
NaOH	Sodium hydroxide
NeP1	Transcriptional repressor negative protein 1 (identical to CTCF)
NFX1-91	Transcriptional repressor NF-X1 isoform 3
NIKS	Immortalised normal keratinocytes
NOTCH1	Notch homolog 1
nt	Nucleotides
Oct1	Octamer-binding transcription factor 1
Oct4	Octamer-binding transcription factor 4
ORF	Open reading frame
P/CAF	K(lysine) acetyltransferase 2B (KAT2B)
p105	Part of the proteasome complex
p107	Retinoblastoma-like 1
p130	Nucleolar phosphoprotein p130
p21	Cyclin-dependent kinase inhibitor 1A
p27	Cyclin-dependent kinase inhibitor 1B
p53	Tumor protein 53
p600	600-kDa retinoblastoma protein- and calmodulin-binding protein
p63	Transformation-related protein 63
PARlation	Poly(ADP)ribosylation
PARP1	Poly(ADP)ribose polymerase 1
pBR322	Vector containing various HPV
pCi	T7 expression vector used for CTCF expression
PCR	Polymerase chain reaction
pDrive	SP6 expression vector for CTCF
PDZ	Structural protein domain often associated with signalling complexes (name derived from Drosophila disc large tumor suppressor (Dlg1) and zonula occludens-1 protein (zo-1))
***PfuUltra* HF**	DNA polymerase for amplification of large template sequences
pGEM2	Vector containing the HPV18 genome
pH	Decimal logarithm of the reciprocal of the hydrogen ion activity in a solution
PIC	Protease inhibitor cocktail
PML	Promyelocytic leukemia protein
pmol	Picomol
PMSF	Phenylmethanesulfonylfluoride
pRb	Retinoblastoma protein
puC19	Vector containing various HPV
PV	Papillomavirus
PWM	Position weight matrix

qPCR	Quantitative real time polymerase chain reaction
Rb	Retinoblastoma
RCF	Relative centrifugal force
RNA	Ribonucleic acid
rpm	Resolutions per minute
RRL	Rabbit reticulocyte lysate
RRP	Respiratory papillomatosis
RV	Reverse
s	Second
SDS	Sodium dodecyl sulfate
SiHa	Squamous cell carcinoma cell line
SMAD	Transcription factor activated by transforming growth factor beta (TGF-β)
Sox9	SRY (sex determining region Y)-box 9
SP1	Specificity protein 1
SP6	Transcription initiation site for the SP6 RNA polymerase
S-phase	DNA synthesis phase of the cell cycle
SSC	Squamous cell carcinoma
SSC13	HPV-negative squamous cell carcinoma line
T4	DNA ligase from bacteriophage T4
T7	RNA polymerase from bacteriophage T7
TBE	Tris-borate-EDTA buffer
TBP	TATA binding protein
TBS/T	Tris-buffered saline and tween 20
TEMED	N,N,N',N'-Tetramethylethylenediamine
Tm	Melting temperature
U	Unit
V	Volt
VEZF1	Vascular endothelial zinc finger 1
W12	Human cervical keratinocyte cell line maintaining HPV16
WGE	Wheat germ extract
WHO	World health organisation
WT	Wild type
YY1	Yin yang 1
$Zn(O_2CCH_3)_2$	Zinc acetate

1 Introduction

1.1 Papillomavirus types and their potential of causing disease

Papillomaviruses (PV) are small epithelium specific DNA viruses that are able to infect a wide range of hosts ranging from reptiles to birds and mammals. The evolutionary origin of these viruses can be traced back to the emergence of amniotes (Garcia-Vallve et al., 2005). As a result of co-evolution with their hosts, papillomavirus infections are very common and often exhibit little or no symptoms (Bottalico et al., 2011, Ekstrom et al., 2010, Ekstrom et al., 2011, Forslund, 2007). However chronic infections with certain papillomavirus types over long periods of time have been shown to drastically increase the risk of certain types of cancer, most notably cervical cancer.

Infections with human papillomaviruses (HPV) are of high prevalence in the general population and some HPV types are able to cause disease ranging from harmless warts to invasive cancer. The standard method of testing for cancer and pre-cancerous lesions in the cervix is the Papanicolaou test, better known as Pap smear or Pap test (Saslow et al., 2012, Papanicolaou, 1948). For this test cell samples are taken and analysed using light microscopy to check for abnormal cell morphology of cervical keratinocytes. Since HPV is a common cause of these abnormalities the test result is also used as an indicator for an HPV infection. To confirm an HPV infection, cell samples can be tested for HPV DNA or RNA (Saslow et al., 2012, Baker, 2013).

There are more than 150 different HPV types identified so far, infecting cutaneous as well as mucosal epithelium (reviewed by Van Doorslaer et al., 2013, Bernard et al., 2010). The worldwide prevalence of HPV infections in sexually active women is between 20 % and 40 % depending on geographical location (Richardson *et al.*, 2003, Clifford *et al.*, 2003). Young women are the part of the population where HPV infections are most common and it is estimated that 80 % of all women acquire a high risk HPV infection during their lives (Baseman and Koutsky, 2005, Bekkers *et al.*, 2004). The majority of these infections are cleared by the immune system within 18 months and re-infection with the same HPV types is considered uncommon (Xi *et al.*, 1995, Bosch *et al.*, 2008, Schiffman *et al.*, 2007, Baseman and Koutsky, 2005, Xie *et al.*, 2007). However, the individuals who fail to clear the infection can maintain the virus over decades. Also clearance is not always 100 % complete and latency of HPV infections has been suggested (Maglennon et al., 2011). The factors that

enable an HPV infection to be maintained over extended periods of time are low antigen presentation on the cell surface, evasion of the innate immune system by inhibiting the interferon response and interference with chemokine signalling for attracting dendritic cells and macrophages (Nees *et al.*, 2001, Perea *et al.*, 2000, Kanodia *et al.*, 2007). Secondary factors that increase the risk of cervical cancer in HPV infected individuals include smoking and the long-term use of oral contraceptives (Haverkos, 2004, Castle, 2004, Moodley, 2004).

The largest proportion of all HPV types (90 %) belongs to the alpha and beta genera. Alpha viruses target mucosal and/or cutaneous epithelium and can be classified as low risk or high risk types according to their association with malignancies (Bernard, 2005, Middleton *et al.*, 2003). Beta papillomaviruses are specialised solely on cutaneous epithelium (Bernard, 2005, Middleton *et al.*, 2003). The low risk types, such as HPV6 and HPV11, are the most common causative agents of genital warts and rarely promote severe disease. However one of the few serious diseases caused by low risk HPV is respiratory papillomatosis (RRP) (Donne and Clarke, 2010). This condition is usually found in juveniles and causes neoplasms in the lung that need to be treated with repeated surgery. Eventually this disease results in lung cancer in 5 % of cases (Derkay, 1995, Hsueh, 2009). However the overall incidence of RRP is low with 4 cases in 100,000 (Gerein *et al.*, 2005, Donne and Clarke, 2010). Low risk HPV infections can also be associated with non-melanoma skin cancer, especially in patients with epidermodysplasia verruciformis (Pisani *et al.*, 2002). The HPV-mediated impairment of DNA damage repair in response to UV light and the inhibition of apoptosis are thought to be the underlying factors for this outcome (Jackson and Storey, 2000, Underbrink *et al.*, 2008). Also the involvement of low risk HPV in some cases of non-melanoma skin cancer has been confirmed (Nindl *et al.*, 2007).

The HPV types classified as high risk types according to the World Health Organisation (WHO) are 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, and 59 (Bouvard *et al.*, 2009, Munoz *et al.*, 2006, Schiffman *et al.*, 2009). A recent meta-analysis using data from around the world revealed that the most common high HPV types are HPV16, 18, 52, 31, and 58 (Bruni *et al.*, 2010). However, the prevalence of a particular HPV type does not always correlate with the degree to which that type contributes to the emergence of invasive cancer. HPV16 and 18 are the most prevalent types and the most common causative agents of HPV induced cancers, but HPV45 is the third most common cancer causing HPV and this virus is not

among the most commonly found HPV (Bruni *et al.*, 2010). Thus there is variation in the carcinogenic potential among different high risk HPV types.

HPV induced carcinomas can be found in a range of tissues. However the most common HPV induced cancer arises at the cervix, in particular in the junction between endo- and exocervix (Herfs *et al.*, 2012). HPV DNA is found in 99.7 % of all cervical carcinomas and an HPV infection is therefore the most important cancer promoting factor in this disease (Walboomers et al., 1999). The carcinomas themselves can be classified again into squamous cell carcinomas (SCC) and adenocarcinomas (AC). HPV16 is the virus that causes 62 % of all SCC of the cervix. HPV18 on the other hand is only connected to 8 % of all cervical SSC. However, type 18 has a more significant contribution to AC since it is involved in the generation of 32 % of all adenocarcinomas in the cervix (de Sanjose *et al.*, 2010). These percentages vary slightly from study to study (Munoz et al., 2003, Munoz et al., 2004, Geraets et al., 2012).

With 530,000 new cases each year worldwide, cervical cancer is the second most common cancer in women after breast cancer (reviewed by de Martel et al., 2012). Accordingly, 0.03 % of the female population are affected by it and about one third of all cervical carcinomas are lethal (Parkin et al., 2005, reviewed by Yugawa and Kiyono, 2009).

The other sites of HPV induced tumours that are commonly found include vulva, vagina, anus, penis and oropharynx including tonsils and base of the tongue (reviewed by Doorbar et al., 2012, zur Hausen, 2009). The site of infection is closely associated with sexual activity. Oral, vaginal and anal intercourse with an infected individual favours an HPV infection at the site of sexual contact (Machalek *et al.*, 2012, Gillison *et al.*, 2012). A prevalence similar to the one of HPV-induced cervical cancer is only observed with HPV-induced anal carcinomas in men who have sex with men (Bosch and de Sanjose, 2003). However, the absolute number of cervical cancer cases outnumbers the amount of HPV induced anal carcinomas greatly (reviewed by de Martel et al., 2012). In contrast to many other sexually transmitted diseases, condoms do not offer sufficient protection against a HPV infection (Cottler *et al.*, 2006, Winer *et al.*, 2006, Baldwin *et al.*, 2004). In a study with a sample size of 82 sexually active female university students the prevalence of genital HPV infection among students who consistently used condoms was found to be 37.8 % compared to 89.3 % in non-condom users (Cottler *et al.*, 2006). However these results are debatable due to the small sample size.

Research in recent years has shown that HPV involvement in cancer of the head and neck causes between 13 % and 56 % of malignancies depending on the region of the world (Forman *et al.*, 2012). This cancer is of high prevalence and affects males three times more often than females (Gillison *et al.*, 2012). The predominant HPV type involved in head and neck cancer is HPV16. This type is covered by the vaccines that are currently used for vaccination against cervical cancer. Thus men would benefit from the vaccination as well but the only part of the population frequently vaccinated in western countries is young women.

1.2 Epidemiology of HPV

Cancer is one of the major health burdens world-wide and a significant proportion of this disease is caused by infectious agents. This opens up the potential to prevent cancer by preventing or treating the underlying infection. From the 12.7 million new cancer cases in 2008, about 2 million were caused by infectious disease as illustrated in Figure 1 (reviewed by de Martel et al., 2012, Parkin, 2011). In turn, 610,000 out of these 2 million cases were caused by HPV resulting in an overall contribution of HPV to the global cancer burden of 4.8 % (reviewed by de Martel et al., 2012). The less developed world has the highest incidence of HPV induced cancer with 490,000 cases. However the proportion of HPV-induced cancer among all infection related cancers is higher in more developed countries.

Figure 2 shows the global incidence of cervical cancer in form of a world map (Forman *et al.*, 2012).

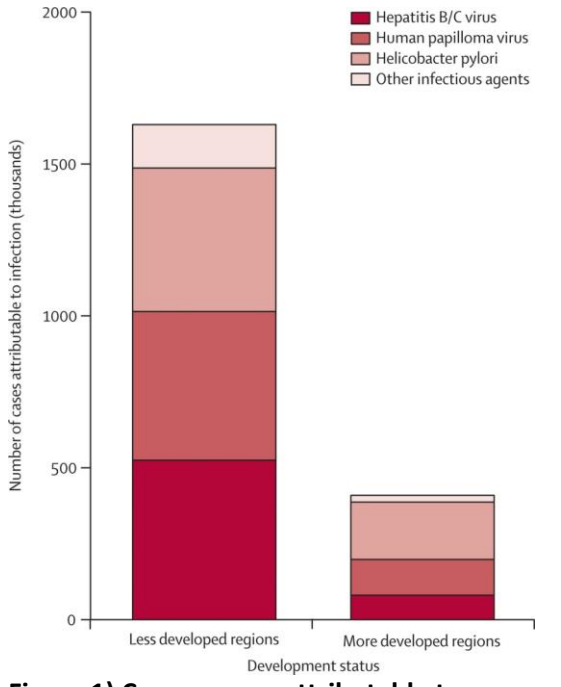


Figure 1) Cancer cases attributable to infection in 2008

HPV is one of the three infectious agents causing the majority of infection related cancer. The other two are Hepatitis B/C and *Helicobacter pylori*. Figure of de Martel *et al.*, reproduced with permission (de Martel et al., 2012).

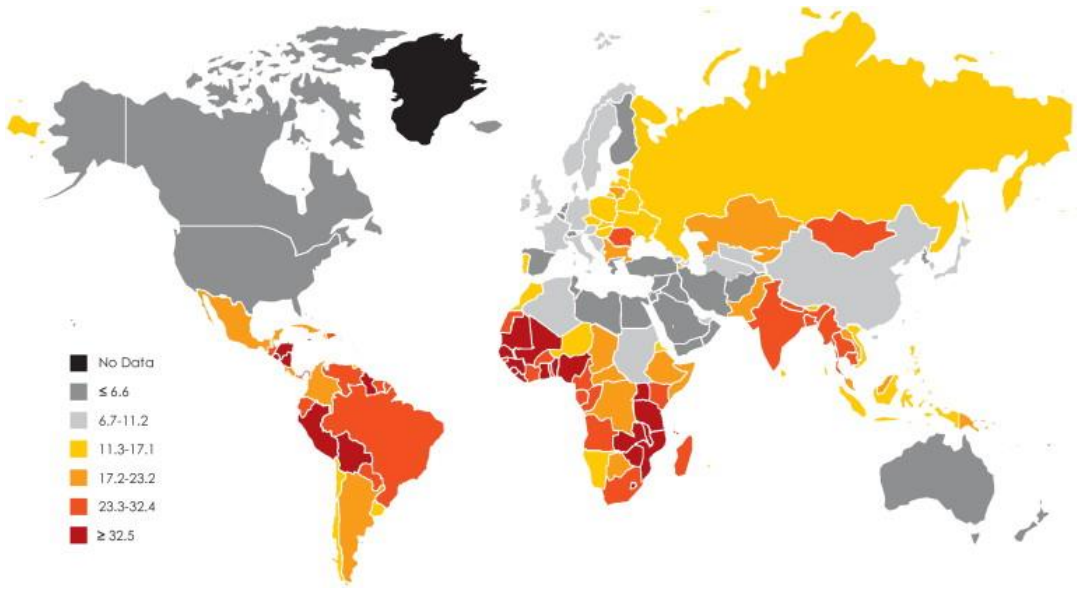


Figure 2) Global incidence of cervical cancer per 100,000
Hot spots for a high incidence of cervical cancer are less developed regions of the world, for example Sub-Saharan Africa and South America. Figure of Forman *et al.*, reproduced with permission (Forman et al., 2012).

Hot spots with the highest incidence with over 30 cases per 100,000 per annum are found in South America and Sub-Saharan Africa. On the other hand the incidence is lowest in North America, Western Europe, North Africa, China and Australia with less than 6.6 cases per 100.000 inhabitants. Parts of the population of every country are carriers of a high risk HPV infection without showing any pathological symptoms (Forman *et al.*, 2012). The percentage of these asymptomatic infections varies greatly among different regions of the world and this percentage does not always correlate to the incidence of cervical cancer as seen in Figure 2 and Figure 3. The incidence of cervical cancer in Eastern Europe is on an intermediate level even though this region has one of the highest percentages of carriers of asymptomatic high risk HPV infections. The reason for a higher

Figure 3) Percentage of asymptomatic HPV infections
Percentage of the female population carrying a high risk HPV infection without showing pathological symptoms. This percentage is highest in the Caribbean and Africa. It mostly correlates with the development level of a country. Figure of Forman *et al.*, reproduced with permission (Forman et al., 2012).

number of cases in some regions can be diverse.

However one of the main factors to consider is the age of the population. Whereas European and North American countries have a relatively high incidence of HPV infections in young women, this incidence declines with increasing age due to clearance of the infection (Franceschi *et al.*, 2006). This decline is not seen in less developed countries. Also the fact that the percentage of infected young women in less developed countries is higher results in an even higher proportion of HPV infected females.

The estimated number of patients per year with HPV-induced cancer from all over the world can be found in Table 1 (reviewed by de Martel et al., 2012). The number of cervical cancer cases dominates the other types of HPV induced cancer with 530,000 cases every year.

Site of tumour	No. of cases	Cancer cases with HPV involvement			
		Percentage of total	Absolute number	In less developed world	In more developed world
Cervix uteri carcinoma	530,000	100 %	530,000	450,000	77,000
Penile carcinoma	22,000	50 %	11,000	7600	3200
Anal carcinoma	27,000	88 %	24,000	12000	12000
Vulva carcinoma	27,000	43 %	12,000	4100	7500
Vaginal carcinoma	13,000	70 %	9,000	5700	3400
Oropharyngeal carcinoma including tonsils and base of tongue	85,000	25.6 %	22,000	6400	15000

Table 1) The contribution of HPV to the global cancer burden in 2008
This is the most recent estimation to date, published in October 2012. Table modified from de Martel et al. (de Martel *et al.*, 2012).

The absolute number of patients with anal cancer is the same in the developed as well as the developing world. The fact that there are many more people living in the developing world than in the developed world indicates that people living in the developed world have a higher chance of developing anal cancer (reviewed by de Martel et al., 2012). This shift towards a higher chance of developing cancer is even more obvious with oropharyngeal carcinomas which already occur twice as much in the developed world despite the lower overall population (reviewed by de Martel et al., 2012).

For the United Kingdom, as a part of the developed world, infection related cancer makes up a smaller proportion of the overall cancer burden compared to the world average of 16. 1 % (Parkin, 2011). The total incidence of infection related cancer in the UK was recently estimated to be 3. 1 % (Parkin, 2011). About half of that, 1.6 %, is attributed to HPV. This percentage is significantly lower than the estimated average of developed countries of 7.4 % (reviewed by de Martel et al., 2012, Parkin, 2011). However, the contribution of HPV in particular cancers in the UK is still significant as seen in Table 2. In total 5088 cases of HPV-induced cancers were recorded in the UK in 2010 (Parkin, 2011). The contribution of HPV in these cancers was determined by the presence of high risk HPV DNA in tumour tissue from all sites with the exception of the oropharynx. Here expression of E6 and E7 needed to be confirmed in addition to the presence of HPV DNA (Parkin, 2011).

Site	Overall cases of cancer		Cases accountable to HPV		
	Male	Female	Male	Female	Proportion of total cases
Cervix uteri	NA	2691	NA	2691	2691 (100 %)
Oral cavity	2284	1421	183	114	296 (8.0 %)
Oropharynx	981	323	138	45	184 (14. 1 %)
Larynx	1803	386	191	40	232 (10.6 %)
Anus	364	621	328	559	887 (90.0 %
Vulva	NA	1128	NA	451	451 (40.0 %)
Vagina	NA	251	NA	157	157 (62.5 %)
Penis	475	NA	190	NA	190 (40.5 %)
Total:	5907	6821	1030	4058	5088 (40 %)

Table 2) Incidence of HPV induced cancer in the United Kingdom in 2010
Table of Parkin *et al.*, reproduced with permission (Parkin, 2011).

1.3 Vaccination

In 2006, the first vaccine against HPV, Gardasil, was approved by the Food and Drug Administration of the USA and the approval for Europe and Australia followed the year after. This vaccine is manufactured by Merck & Co. and immunises against four HPV types, the high risk types HPV16/18 and the low risk types HPV6/11 (Siddiqui and Perry, 2006). In 2007 a second HPV vaccine, Cervarix, was approved in the European Union which immunises against the high risk types HPV16/18 only (Crosbie and Kitchener, 2007). The manufacturer of this vaccine is GlaxoSmithKline. Both vaccines have been shown to be safe with only minor side effects (Bayas *et al.*, 2008). Klein *et al.* conducted a study in which the side effects of the quadrivalent HPV vaccine with a sample size of 189,629 females were assessed (Klein *et al.*, 2012). They concluded that the only side effects found were occasional syncope on the day of vaccination and a slightly increased chance of acquiring a skin infection within two weeks following vaccination (Klein *et al.*, 2012).

Virus-like particles containing the HPV protein L1 are the basis for both vaccines (Crosbie and Kitchener, 2007, Siddiqui and Perry, 2006). This protein is only found in mature viral particles and the uppermost layer of infected epithelium. Since the majority of cells of already infected epithelium do not express L1, those cells are not affected by the immune response induced by the vaccine. Thus the vaccine is thought to only protect from new infections caused by L1-coated mature viral particles (Markowitz, 2011). Immunisation against the two most common high risk types before the first contact with the virus is predicted to result in the reduction of the overall risk of developing cervical cancer by about 70 % since the two high risk types covered by the vaccine cause 70 % of cervical cancer (reviewed by Yugawa and Kiyono, 2009, de Sanjose *et al.*, 2010). Considering that there is a degree of antibody cross-reactivity to other HPV types, the remaining risk of acquiring HPV-associated cervical carcinoma may in fact be lower than 30 % (Harper *et al.*, 2006, Cutts *et al.*, 2007, Paavonen *et al.*, 2007). However HPV infections from HPV types not covered by the vaccine, for example HPV45 and HPV31, will continue to cause cervical cancer (Paavonen *et al.*, 2009). Hence regular pap smears are still being recommended to vaccinated women (Siddiqui and Perry, 2006).

In addition to the protection against two high risk types, Gardasil also grants protection against HPV6 and HPV11 which are responsible for more than 80 % of all genital warts (Schmiedeskamp and Kockler, 2006, Winer *et al.*, 2005). These warts are found in about 1 %

of sexually active adults and impose a £16.8 million burden on the NHS annually from cases in England only (Koutsky, 1997, Desai *et al.*, 2011).

Cervarix was shown to have a good long-term safety profile (Harper *et al.*, 2006). More than 98 % seropositivity for HPV16/18 antibodies was observed 4.5 years after vaccination (Harper *et al.*, 2006). Thus it is expected that no follow-up vaccination is needed. In a study conducted over 4 years with more than 10,000 females, Lethinen *et al.* confirmed a 100 % protection against high grade cervical intraepithelial neoplasm and adenocarcinoma *in situ* caused by HPV16 and 18 in vaccinated individuals who had received the HPV vaccine prior to their first contact with the virus (Lehtinen *et al.*, 2012). Additionally, it is worth noting that this study was funded by one of the vaccine manufacturers, GlaxoSmithKline. Although a review of various trials for both vaccines confirmed the efficacy of the vaccines against HPV infection by the HPV types covered (Schiller *et al.*, 2012). No therapeutic efficacy of the vaccine in already infected individuals could be confirmed (Schiller *et al.*, 2012).

In most countries with a HPV vaccination program only females are vaccinated. However due to the existing evidence for HPV-induced cancer in other tissues, and the transmission via males, it would be beneficial to extend the vaccination programme to the general population to reduce the overall incidence of HPV-induced cancer. The chance of a male developing HPV-induced cancer is lower compared to a female so vaccinating males comes down to a matter of cost (Russell *et al.*, 2013). Males are mainly perceived as carriers of HPV who less often progress to severe disease and therefore do not qualify for the expense needed for vaccination.

The search for HPV vaccines targeting a wider spectrum of HPV types continues and a phase 3 clinical trial of a third HPV vaccine called V503 (Merck) will be completed in 2013. If this vaccine should become available it will cover 9 HPV types and therefore provide a higher degree of protection than currently available vaccines. The 9 HPV types covered cause 87 % of cervical cancer (de Sanjose et al., 2010). In conclusion, vaccination will reduce the incidence of HPV induced cancer in the general population but the need for further research remains due to the limitations of the vaccine in terms of type specificity, the restriction to prophylactic use and high vaccination costs.

1.4 Papillomavirus genetics and transcriptional regulation

The HPV genome (Figure 4) consists of approximately 8kb of double stranded, circular DNA and encodes 6 early and 2 late open reading frames (ORFs). The early ORFs encode for E1, E2, E1^E4, E5, E6 and E7 whereas the late ORFs encode for L1 and L2.

The majority of viral gene expression as well as the replication of the viral genome is controlled by a single long control region (LCR) (reviewed by Graham, 2010). This LCR contains the origin of replication as well as binding sites for various transcription factors including the viral regulatory protein E2, the viral helicase E1, the host transcription factors SP1, TBP, MYC and others (Steger and Corbach, 1997, Sichero *et al.*, 2012). Another feature of the LCR is an oestrogen-triggered hormone response element that was shown to be involved in carcinogenesis (Gariglio *et al.*, 2009, Arbeit *et al.*, 1996).

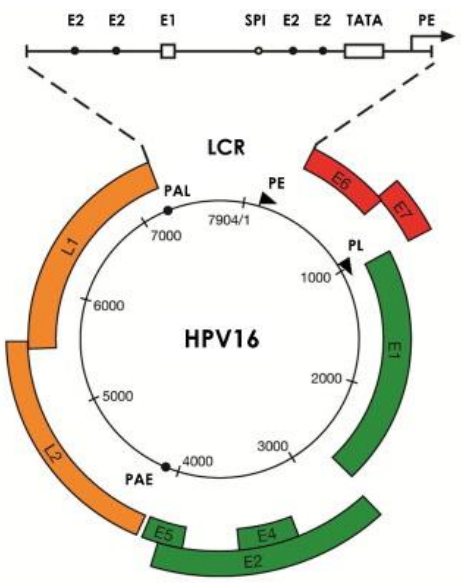


Figure 4) The HPV16 genome
Early genes are shown in red (oncoproteins) and green (other early proteins). Late genes are coloured orange. All these genes are regulated via the long control region (LCR). Figure from Doorbar *et al.*, reproduced with permission (Doorbar et al., 2012).

Methylation of transcription factor binding sites in the LCR plays a major role in viral gene regulation. This methylation pattern changes when the host cell differentiates or when it progresses through the different stages of carcinogenesis (reviewed by Johannsen and Lambert, 2013). Methylation in the LCR mostly affects the major viral transcription factor E2. Dimers of this protein recognise and bind palindromic DNA sequences (AACCg(N4)cGGTT) and regulate gene expression throughout all stages of the viral life cycle (Dell *et al.*, 2003). Methylation of E2 binding site abrogates E2 binding and methylation of particular E2 binding sites has been linked to hyperactive transcription of viral oncogenes (Vinokurova and von Knebel Doeberitz, 2011, Kim *et al.*, 2003, Thain *et al.*, 1996).

In the early stages of the viral life cycle, low levels of E2 enhance transcription of the early genes (Steger and Corbach, 1997). This action is reversed at the later stages of the viral life cycle where E2 protein accumulates and triggers downregulation of the viral oncogenes E6

and E7 by displacing SP1 from the LCR, causing proliferation arrest while promoting differentiation of the host cell (Steger and Corbach, 1997).

Next to its role in gene regulation E2 is an important factor in the amplification of the viral genome since it has been shown to recruit E1 to the origin of replication (Masterson *et al.*, 1998). This mechanism is important for the initial amplification of viral genomes upon infection and for the amplification of viral genomes for the assembly of viral particles (Mohr *et al.*, 1990, reviewed by Doorbar et al., 2012). However, it is suggested that E2 can ensure maintenance of viral genomes independently of E1 (Egawa *et al.*, 2012).

In all papillomaviruses at least one inhibitory variant of E2 exists (reviewed by McBride, 2013). This variant includes a part of the otherwise untranslated E8 ORF on its N-terminus and is therefore called E8^E2 (Hubbert *et al.*, 1988). E8^E2 competes for E2 binding sites and can form heterodimers with full length E2. These heterodimers are negative regulators of transcription and are insufficient for long term maintenance of HPV episomes (Kurg *et al.*, 2010, Stubenrauch *et al.*, 2000).

In papillomavirus genetics the position given in the promoter name refers to the actual start site of the mRNA regulated by this promoter. The promoter sequences of the two major promoters of PV are situated in the LCR but the only mRNA start site in the LCR belongs to the early promoter (PE) (reviewed by Van Doorslaer et al., 2013). In HPV16 the start site of the early promoter is located at nucleotide 97 and the transcription start site of the late promoter is at nucleotide 670 as seen in Figure 4.

Other important features of HPV genomes are the two polyadenylation sites which are named early and late polyadenylation site. Promoter choice, splicing and the choice of the polyadenylation site determine which transcript is produced (reviewed by Graham, 2010, Florin et al., 2002, Johansson et al., 2012). Most HPV mRNAs are polycistronic, thus splicing is crucial (Stacey *et al.*, 2000).

Transcripts from the early promoter using the early polyadenylation signal can encode for the viral oncoproteins E6 and E7 (Stacey *et al.*, 2000). These proteins immortalise the host cell, drive cell proliferation and inhibit apoptosis (reviewed by Yugawa and Kiyono, 2009).

Transcripts from the late promoter using the early polyadenylation site can encode for the proteins needed to amplify the viral genome. These are E1, E2, E1^E4 and E5 (Bodily and Laimins, 2011). On the other hand transcripts from the late promoter that use the late polyadenylation signal encode for the proteins necessary for virus assembly and release of

mature viruses (Day *et al.*, 1998, Wang *et al.*, 2004). These proteins are E1^E4, L2 and L1 (Bodily and Laimins, 2011).

More specific gene regulation from either of these polycistronic transcripts is achieved by splicing. For example the splice variant encoding L2 is favoured first until enough L2 is generated. Subsequently the splice variant encoding L1 is favoured (Florin *et al.*, 2002).

1.5 Papillomavirus life cycle

The entry point of a papillomavirus infection is suggested to be a microlesion in the host's epithelium that enables the virus to reach the basal epithelial cells (Schiller *et al.*, 2010). Upon infection of a basal cell, the viral DNA migrates into the nucleus where the HPV genome is initially amplified prior to being maintained as a stable episome of about 200 copies per cell (Pyeon *et al.*, 2009, reviewed by McBride, 2008). With each cell division, the amount of viral episomes is split equally between both daughter cells. An E2-mediated tethering mechanism to host chromosomes involving the host helicase ChlR1 is crucial for this event and also insures correct partitioning of viral episomes to daughter cells (McBride *et al.*, 2006, Van Tine *et al.*, 2004a, Dao *et al.*, 2006, Parish *et al.*, 2006, reviewed by McBride, 2008). One of the daughter cells resides in the basal layer whereas the second daughter cell migrates towards the epithelial surface, differentiating on its way (reviewed by Herrington, 2009). This differentiation process triggers viral maturation and results in the production of infective viral particles which are shed off at the surface of the epithelium. Hence an HPV infection is not accompanied by premature host cell death or cell lysis, so release of inflammatory cytokines or viremia do not occur (Stanley, 2006). Healthy cells only divide in the basal layer of the epithelium whereas HPV infected cells also undergo mitosis in the suprabasal layer (reviewed by Doorbar et al., 2012). In both layers the viral oncogenes E6 and E7 are expressed which promotes cell proliferation while inhibiting apoptosis. They push the cells in the suprabasal layer into an S-like phase followed by a G2-like phase (Banerjee *et al.*, 2011). However particular functions of these early proteins differ in high risk and low risk HPV (reviewed by Klingelhutz and Roman, 2012).

In the mid-epithelial layers the amplification of viral episomes takes place (reviewed by Klingelhutz and Roman, 2012, Felsani *et al.*, 2006). To facilitate this, the viral proteins E1, E2, E1^E4 and E5 are expressed. E2 is needed to recruit the viral helicase E1 to the origin of

replication which is crucial for initial amplification of the genome and enabling the virus to replicate efficiently in the later stages of the viral life cycle (Bodily and Laimins, 2011, Kim and Lambert, 2002, Angeletti *et al.*, 2002). E1 may enable the viral episome to be replicated independently from the host genome leading to an increase in viral genomes of at least 2 log in animal models (Banerjee *et al.*, 2011, Wang *et al.*, 2009, Blakaj *et al.*, 2009, reviewed by Doorbar, 2006). Despite its role in viral genome amplification at advanced stages of the life cycle, E1 is suggested to be dispensable for episome maintenance in basal cells (Egawa *et al.*, 2012). The E5 protein is involved in enhancing genome amplification by promoting koilocyte formation and activating the nuclear localisation signal of E1 which is in turn needed for genome amplification (Crusius *et al.*, 2000, Crusius *et al.*, 1998, Krawczyk *et al.*, 2008). E5 is a pore-forming transmembrane protein with a cytoplasmic C-terminus and is known to interfere with apoptosis and the trafficking of vesicles (Krawczyk *et al.*, 2011, Kabsch *et al.*, 2004). Another protein involved in the amplification of viral genomes is E1^E4. The abundant expression of E1^E4 causes growth arrest of the host cell and aides amplification of viral genomes (Wilson *et al.*, 2005, Wilson *et al.*, 2007, Peh *et al.*, 2004).

In one of the uppermost layers of the epithelium, the granular layer, the late genes L2 and L1 are expressed in consecutive order. When the cells of the granular layer exit the cell cycle the minor capsid protein L2 is expressed (Day *et al.*, 1998). L2 is recruited to the already amplified viral genomes prior to the expression of L1. When the major capsid protein L1 is expressed it assembles with the complex of viral DNA and L2 to form a complete early virion (Day *et al.*, 1998, Holmgren *et al.*, 2005). The absence of mitochondrial activity in the granular layer of the epithelium causes a switch in the environmental condition from a reducing environment to an oxidising environment, thus enabling the formation of disulphide bonds between L1 protein resulting in the completion of the viral particle (Buck *et al.*, 2005, Finnen *et al.*, 2003). The mature viral particle consists of pentameric capsomeres which are formed by 360 L1 proteins and a smaller and variable number of L2 proteins (Buck *et al.*, 2008).

Additionally, E1^E4 is also expressed in the uppermost layer of the epithelium which is thought to have a role in the release of mature viral particles by disrupting the keratinous layer of the skin due to amyloid formation (McIntosh *et al.*, 2008, Doorbar *et al.*, 1997). Accordingly, loss of E1^E4 interferes with late events in the viral life cycle (Wilson *et al.*, 2005, Nakahara *et al.*, 2005). The viral gene expression in each layer of the epithelium is illustrated in Figure 5.

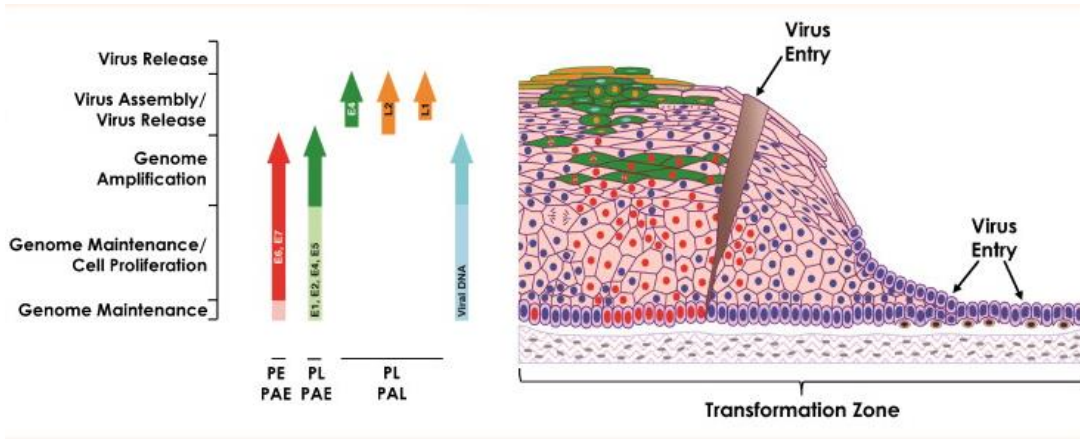


Figure 5) HPV gene expression in mucosal epithelium
A microlesion is shown that serves as point of entry for an HPV infection. HPV infected, actively dividing cells are marked with a red nucleus. Viral gene expression is closely connected to the differentiation state of the host cell. In different layers of the skin, all cells are at a defined stage of the differentiation process. The early genes are expressed in the basal and suprabasal layers and induce mitosis in these layers. In the mid layers of the epithelium E1, E2, E1^E4 and E5 are expressed, each playing a role in the amplification of the genome. In the top layers the late genes L1 and L2 are expressed simultaneously with E1^E4. These proteins are needed for virus assembly and release. HPV can also infect cells adjacent to the epithelium in the squamocolumnar junction of the cervix. Here the viral life cycle may be inefficiently supported but the risk for developing cancer as a result of an HPV infection is high (Herfs et al., 2012). Figure from Doorbar *et al.,* reproduced with permission (Doorbar et al., 2012).

1.6 Mechanisms of HPV induced carcinogenesis

1.6.1 Progression from healthy epithelium to invasive cancer

HPV can infect various kinds of epithelium; however HPV-induced tumours most commonly arise in the cervix, in particular at the junction between endo- and exocervix (Herfs *et al.*, 2012). Progression through the precancerous states to a tumour mostly occurs if a high risk HPV infection is maintained over many years or decades. The average time of persistence differs with HPV type and HPV16 has an especially long time of persistence (Koshiol *et al.*, 2006, Schiffman *et al.*, 2010).

Efficient evasion of the immune system is the key to long term persistence of an HPV infection. Thus HPV has evolved strategies to evade the host immune system in order to prevent clearance of the infection. These include limited gene expression in the basal layer of the epithelium which results in low immunogenicity (Hibma, 2012, Stanley, 2009). In

suprabasal layers expression of E5-results in the degradation of MHC class I molecules which leads to decreased display of viral peptides on the cell surface (Campo *et al.*, 2010). Also the viral oncoproteins E6 and E7 have been shown to play a role in evasion of the immune system as they suppress the interferon response and downregulate other cytokines such as interleukins (Stanley, 2012, Kanodia *et al.*, 2007, Karim *et al.*, 2011).

The most common cancer occurring in the cervix is squamous cell carcinoma (SCC), a tumour of cutaneous or mucosal epithelium. Various HPV types are involved in the emergence of SSC and the contribution of particular HPV types to SCC varies. The majority of SCC is caused by HPV16 (62 %), 18 (8 %), 45 (5 %), 31 (4 %) and 33 (4 %) (de Sanjose *et al.*, 2010). There are more HPV types associated with SCC but their contribution to the overall cases of SCC is below 4 % (de Sanjose *et al.*, 2010).

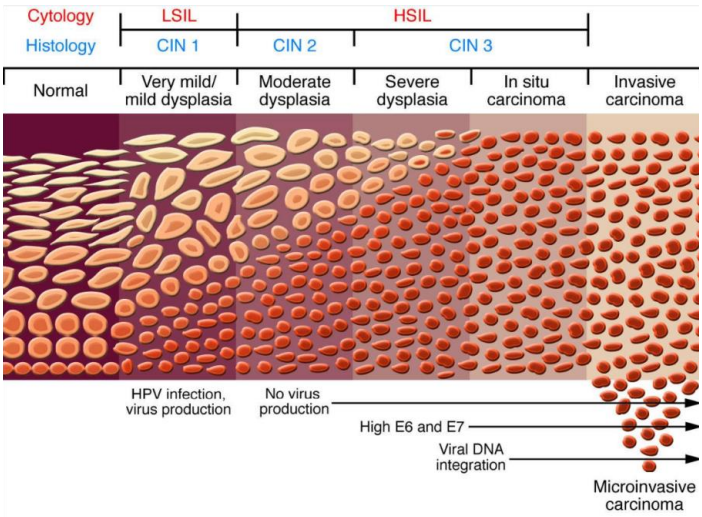


Figure 6) Histology of HPV induced carcinogenesis
Histology of HPV infected epithelium through the pre-cancerous stages to invasive carcinoma. In mild dysplasia the virus is able to complete its whole life cycle. This ability is lost with further progression of the dysplasia. The expression of the viral oncogenes E6 and E7 usually increases during the progression and the final emergence of a tumour often involves semi-random integration of the viral episome into the host genome. Figure from Lowry and Schiller, reproduced with permission (Lowy and Schiller, 2006).

Another kind of malignancy commonly caused by HPV is adenocarcinoma (AC), a tumour of glandular origin. Here the group of HPV types causing this cancer includes less HPV types. The types associated with 94 % of AC are HPV16 (50 %), 18 (32 %) and 45 (12 %) (de Sanjose *et al.*, 2010). All other HPV types have a contribution to AC of less than 1 % (de Sanjose *et al.*, 2010). There are two different nomenclatures frequently used for describing the different histological stages caused by an HPV infection that ultimately lead to invasive carcinoma. These stages are illustrated in Figure 6. The first nomenclature uses the acronyms LSIL and HSIL, referring to a low or high grade squamous intraepithelial lesion. A mild epithelial dysplasia is referred to as LSIL whereas moderate and severe

dysplasia together with carcinoma *in situ* are referred to as HSIL. The second nomenclature for describing the severity of HPV induced lesions includes the terms cervical intraepithelial neoplasm of the grades 1 to 3 (CIN1-3). CIN1 refers to mild dysplasia, CIN2 to moderate dysplasia and the term CIN3 covers severe dysplasia and carcinoma *in situ.* Progression past HSIL or CIN3 results in invasive carcinoma (Lowy and Schiller, 2006). However this only happens in the minority of cases. Typically, progression through the stages of carcinogenesis takes many years and the stages CIN1 and CIN2 are completely reversible to healthy epithelium.

Figure 7 illustrates the time course of HPV infections. Even though high grade dysplasia is mostly seen when the infection is maintained over decades, occasionally young women progress to CIN2+ shortly after infection (Paavonen *et al.*, 2007, Paavonen *et al.*, 2009, Szarewski *et al.*, 2012, Quint *et al.*, 2012). This is suggested to be caused by hormonal changes or epigenetic modification of the HPV DNA (Peh *et al.*, 2004).

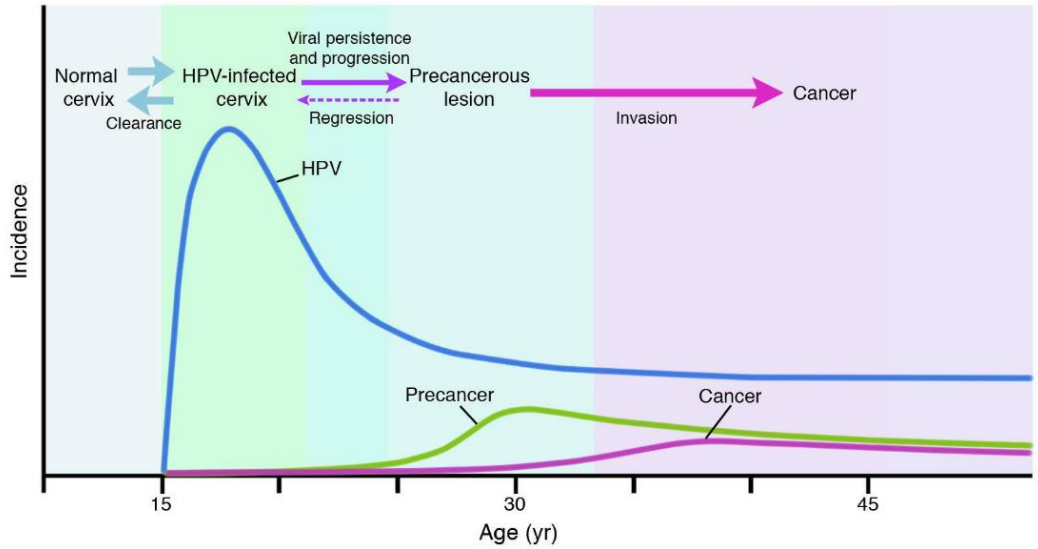


Figure 7) Time course of high risk HPV infections
The pre-cancerous histological stages CIN1 and CIN2 are reversible. However if the pre-cancerous stages are maintained over many years the likelihood of progression to invasive cancer increases. Figure from Lowy and Schiller, reproduced with permission (Lowy and Schiller, 2006).

Carcinogenesis is a process involving many molecular changes in the cell such as inhibition of apoptosis pathways, upregulation of growth factors and immortalisation of the cell. All these changes need to accumulate in a single cell to enable it to uncontrollably proliferate. The papillomavirus oncoproteins E6 and E7 facilitate many of these changes as shown in Figure 8 and Figure 9. Thus, only minor additional alterations are needed to push HPV

infected cells into a cancerous state. When dysplasia increases in severity, typically the amount of E6 and E7 expressed increases (Melsheimer *et al.*, 2004).

This in turn goes along with decreased host genome stability and an increased chance of integration of the viral genome into the host genome at common fragile sites (Van Tine *et al.*, 2004b, Ziegert *et al.*, 2003, Thorland *et al.*, 2003). However the cancer promoting functions of E6 and E7 vary among HPV types. Accordingly, the oncoproteins of high risk HPV are more potent in promoting cancer compared to the ones from low risk HPV (White *et al.*, 2012). In the majority of carcinoma *in situ* and invasive carcinomas of the cervix the HPV genome was found to be integrated into the host genome (Badaracco *et al.*, 2002). Whereas HPV18 DNA is integrated in nearly all HPV18 induced carcinomas, around 30 % of all HPV16 induced carcinomas show no integration of the HPV16 genome (reviewed by Doorbar et al., 2012, Badaracco *et al.*, 2002, Woodman *et al.*, 2003). However, this percentage varies greatly from study to study (Pett and Coleman, 2007, Vinokurova *et al.*, 2008, Matsukura *et al.*, 1989). Integration usually happens after the HPV infection was maintained over an extended period of time. In this case the viral oncogenes are often already de-regulated in the episomes and the integration event de-regulates these oncogenes even further (Melsheimer *et al.*, 2004).

The way the viral genome integrates into the host genome is crucial for progression to cancer. Integrated viral DNA must contain the LCR and the early genes E6 and E7 in addition to the 5' portion of the E1 gene, but the E2 reading frame must be either disturbed or silenced (Pett and Coleman, 2007). Since E2 is the negative regulator protein of the viral oncogenes, the lack of E2 leads to overexpression of these genes (Jeon *et al.*, 1995, Jeon and Lambert, 1995, Pett and Coleman, 2007). Also, E2 is able to inhibit E6 and E7 function by direct protein-protein interactions (Gammoh *et al.*, 2006, Grm *et al.*, 2005, Smal *et al.*, 2009). Thus loss of E2 expression is an important factor in HPV-induced carcinogenesis. Most viral genomes found in tumours integrated within the E1 or L1 reading frame (Xu *et al.*, 2013, Theelen *et al.*, 2013, Wang *et al.*, 2013, Dall *et al.*, 2008). Integration in E1 results in loss of E2 whereas integration in L1 leaves the E2 reading frame intact which makes silencing of E2 necessary for cancer progression.

1.6.2 Functions of high risk HPV oncoproteins

The differences in the viral oncoproteins E6 and E7 among various HPV types determine the carcinogenic potential of each HPV type (Pim and Banks, 2010, reviewed by Klingelhutz and Roman, 2012). HPV induced carcinogenesis works via multiple molecular mechanisms, one of which is the ubiquitin dependent degradation of tumour suppressor genes. Ubiquitin is a small protein that can be attached to target proteins as a posttranslational modification called ubiquitination. This modification can influence protein localisation, activity or mediate the degradation of the target protein by the proteasome (reviewed by Glickman and Ciechanover, 2002, reviewed by Mukhopadhyay and Riezman, 2007, Shih et al., 2000). A single ubiquitin or chains of ubiquitin can be attached to lysines or the N-terminus of the target protein in a three-step process consisting of activation, conjugation and ligation (Bloom et al., 2003, reviewed by Pickart and Eddins, 2004, Flick et al., 2004). Firstly, an ubiquitin-activating enzyme (E1) catalyses the acyl-adenylation of the C-terminus of the ubiquitin molecule (reviewed by Schulman and Harper, 2009, reviewed by Komander and Rape, 2012). This results in a thioester linkage between the C-terminal carboxyl group of ubiquitin and the E1 cysteine sulfhydryl group. Secondly, an ubiquitin-conjugation enzyme mediates the transfer of ubiquitin from the activating enzyme to a cysteine of the conjugation enzyme (reviewed by van Wijk and Timmers, 2010). Thirdly an ubiquitin ligase forms an isopeptide bond between the C-terminal glycine of the ubiquitin and a lysine of the target protein (reviewed by Metzger et al., 2012). Ubiquitin ligases can have two different domains that can catalyse the reaction. The homologous to the E6AP carboxyl terminus (HECT) domain binds ubiquitin transiently whereas the really interesting new gene (RING) domain mediates the transfer of ubiquitin from the conjugation enzyme to the target protein (reviewed by Metzger et al., 2012). A chain of at least 4 ubiquitin proteins attached by lysine 48 to one another mediates degradation by the proteasome, cutting the target protein into small polypeptides (Thrower et al., 2000).

E6 is able to form complexes with the ubiquitin ligase E6-associated protein (E6AP) and alter its specificity (Gewin *et al.*, 2004, Tomaic *et al.*, 2009). In this mechanism E6 seems to work like an adapter protein for E6AP, enabling E6AP to target various tumour suppressor genes like p53, ΔN63α and others, facilitating their degradation by the proteasome (Figure 8) (Xu *et al.*, 2008, Sherman *et al.*, 2002, Senoo *et al.*, 2007). Degradation of the major tumour suppressor protein p53 through this pathway is considered as the most important carcinogenic function of E6. Other proteins that are degraded include ΔN63α which is a

splice variant of p63 that plays a role in the regulation of cell proliferation. Its degradation favours cell proliferation (Patturajan *et al.*, 2002).

E6 is also able to target host PDZ domain proteins for degradation via its PDZ binding domain. This function was shown to be critical for the transformation ability of E6 (Kiyono *et al.*, 1997, Watson *et al.*, 2003). Proteins deactivated due to their PDZ domain include various tumour suppressor proteins like hDLG, hScribble, MUPP1, MAGI-1 and others. Particularly the downregulation of the tight junction protein MUPP1 contributes to anchorage independency in cancer development (Kiyono *et al.*, 1997, Lee *et al.*, 1997, Nakagawa and Huibregtse, 2000, Lee *et al.*, 2000). Additionally E6 is able to induce and activate the catalytic subunit of human telomerase, hTERT. This, in combination with the deactivation of the Rb pathways mediated by E7, is sufficient to immortalise cells (Kiyono *et al.*, 1998, Haga *et al.*, 2007). E6 mediated hTERT activation was shown to work via ubiquitination and degradation of a newly discovered transcriptional repressor of the hTERT promoter called NFX1-91 (Xu *et al.*, 2008). Removal of NFX1-91 allows c-myc binding to the hTERT-promoter resulting in induction of hTERT expression. Furthermore E6 upregulates c-myc expression which enhances hTERT expression even further (McMurray and McCance, 2003).

The downregulation of pro-inflammatory cytokines like type I interferons and various interleukins is another cancer promoting function of E6 and E7 since this contributes to the

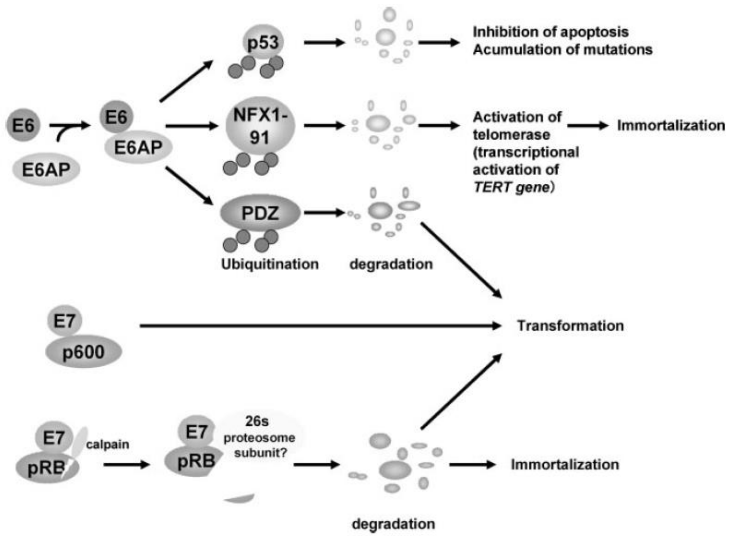


Figure 8) Cancer promoting functions of high risk HPV E6 and E7

E6 and E7 mediated degradation of cellular factors that contribute to tumour formation. E6 modifies the specificity of the host ubiquitin ligase E6-AP so that it ubiquitinates p53, NFX1-91 and proteins containing PDZ domains. This results in the degradation of these proteins and therefore promotes cancer formation. E7 binds and deactivates p600, resulting in reduced anchorage dependency. Additionally E7 mediates calpain-dependent degradation of pRb, enabling the cell cycle to continue in presence of DNA damage. Figure from Yugawa and Kiyono, reproduced with permission (Yugawa and Kiyono, 2009).

ability to evade the immune system (Kanodia *et al.*, 2007, Karim *et al.*, 2011). The lack of pro-inflammatory cytokines prevents the recruitment of Langerhans cells and dendritic cells to the site of infection (Stanley, 2012).

The inhibition of the differentiation of basal cells into keratinocytes was also found to be critical for the formation of a tumour. Differentiation of epithelial cells in tissue culture can be induced by exposure to serum and calcium and it was shown that E6 is able to suppress the differentiation induced by these stimuli (Sherman *et al.*, 2002). Another important factor in the differentiation of keratinocytes is the Notch homolog 1 (NOTCH1) pathway (Nicolas *et al.*, 2003, Proweller *et al.*, 2006). The downregulation of this pathway is important for sustained E6 and E7 expression and tumour development in late CIN stages (Talora *et al.*, 2002). Ectopic expression of NOTCH1 in the HPV-transformed cell lines HeLa and CaSki caused growth suppression, providing further evidence of a role of this pathway in cervical carcinogenesis (Talora *et al.*, 2002). However, publications about NOTCH1 are partly contradictive, and it has also been shown that enhanced NOTCH1 expression can promote cervical carcinogenesis (Veeraraghavalu *et al.*, 2005, Rangarajan *et al.*, 2001, Nair *et al.*, 2003).

The second viral oncoprotein, E7, contains two conserved regions (CR1 and CR2) at its N-terminus, which serve various cancer promoting functions. The domain CR1 plays a role in transformation and has been shown to bind and deactivate p600 and loss of p600 activity induces anchorage independency (Huh *et al.*, 2005, DeMasi *et al.*, 2005). Also the proteins p105, p107 and p130 are degraded by E7, enabling the cell to re-enter the cell cycle in the suprabasal layer of the epithelium (Barrow-Laing *et al.*, 2010, Roman, 2006). Furthermore, CR1 interacts with P/CAF and reduces its acetyltransferase ability (Avvakumov *et al.*, 2003). Acetylation of pRb was shown to be needed for exiting the cell cycle and the acetylation of pRb is lowered upon P/CAF deactivation (Nguyen *et al.*, 2004, Nguyen and McCance, 2005). However this is not the only way for E7 to deactivate pRb. Via CR2, high risk E7 is able to bind pRb and promote degradation by the calcium activated cysteine protease calpain (Darnell *et al.*, 2007). Retinoblastoma protein prevents cell cycle progression when DNA damage is detected. The degradation of this protein combined with the accumulation of DNA damage seen in high grade CIN contributes to cancer development.

Also epigenetic changes are thought to be caused by E7 mediated upregulation of de-methylating enzymes like KDM6A and KDM6B to alter the gene expression of the host cell (McLaughlin-Drubin *et al.*, 2011, Laurson *et al.*, 2010, Hyland *et al.*, 2011). However the

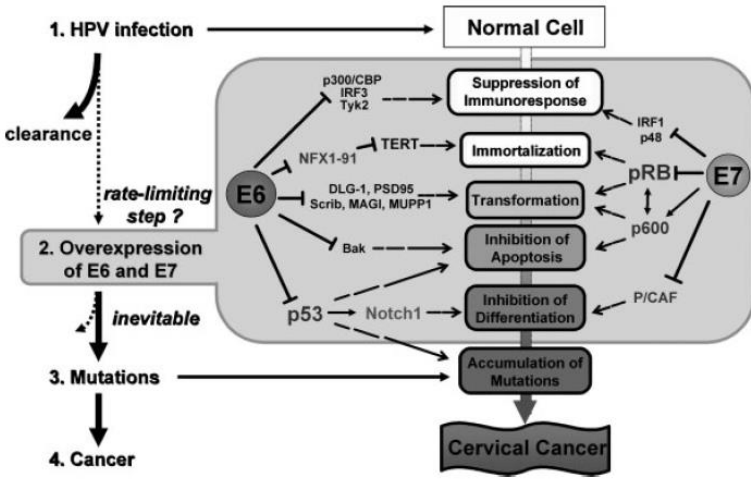


Figure 9) Cellular level of HPV induced carcinogenesis
The contribution of the viral oncoproteins in the acquisition of cancer specific traits is shown. The possible clearance of infection by the host immune system is shown on the left side of the figure. Figure of Yugawa and Kiyono, reproduced with permission (Yugawa and Kiyono, 2009).

alteration of the methylation pattern caused by an HPV infection is diverse since the host methyltransferase DNMT1 was also shown to be overexpressed in infected human foreskin keratinocytes (HFK) as well as in cervical carcinomas (Au Yeung *et al.*, 2010, Leonard *et al.*, 2012). However this does not only affect the host DNA since the methylation pattern of the HPV LCR changes when a lesion progresses through the stages of CIN (Vinokurova and von Knebel Doeberitz, 2011).

Other functions of E7 include the upregulation of the proto-oncogene DEK (and others) and the circumvention of promyelocytic leukaemia protein (PML)-induced senescence (Wise-Draper *et al.*, 2005, Wise-Draper *et al.*, 2006, Bischof *et al.*, 2005). The mechanism how the viral oncoproteins promote carcinogenesis is summarised in Figure 9.

Nevertheless the host cell is not defenceless towards E7 since it has mechanisms to inhibit E7. The host proteins p21 and p27 are able to bind E7 with the effect of impairing its function (Noya *et al.*, 2001). Accordingly, cells containing high levels of p21 and p27 are not pushed into S-phase via E7, in contrast to cells containing low levels of p21 and p27 (Noya *et al.*, 2001).

1.6.3 Functions of E6 and E7 from low risk HPV

The overall incidence of cancer caused by low risk HPV types is substantially lower compared to the cancers caused by high risk HPV types. The reasons for this mostly lie within the functional differences of the early proteins E6 and E7. The low risk E6 protein also binds p53 but does not target it for degradation (Zanier *et al.*, 2012). Also the binding of PDZ-domain proteins and the degradation of those proteins does not occur with low risk E6 protein (Fu *et al.*, 2010, Pim and Banks, 2010).

Furthermore the oncogenic functions of low risk E7 are significantly reduced. The binding of pRb, p105 and p107 is weaker and no degradation of those proteins takes place (reviewed by Klingelhutz and Roman, 2012, reviewed by Doorbar et al., 2012). However low risk E7 is still able to bind and degrade p130 and thus enables re-entry in the cell cycle in mid-epithelial layers (reviewed by Klingelhutz and Roman, 2012, Roman, 2006).

1.7 The host protein CTCF – properties and functions

The CCCTC binding factor, also known as CTCF, is an essential eukaryotic protein with a large variety of functions (Heath *et al.*, 2008). It was first discovered as a transcriptional repressor of the chicken Myc and lysozyme genes (Baniahmad *et al.*, 1990, Lobanenkov *et al.*, 1990). At this time CTCF was also called transcriptional repressor negative protein 1 (NeP1) until it was proven that CTCF and NeP1 were identical (Burcin *et al.*, 1997). CTCF is most known for being a genomic insulator and its binding sites are often located at either side of a gene cluster where CTCF separates heterochromatin from euchromatin (Palstra *et al.*, 2003, reviewed by Wallace and Felsenfeld, 2007). In this position CTCF prevents the spread of heterochromatin into euchromatin, thus maintaining transcriptionally active and inactive regions (reviewed by Ohlsson et al., 2010, Probst *et al.*, 2009). Also, CTCF was found to bind in the spacer region between nucleosomes and it was suggested that it might play a role in nucleosome positioning (Kanduri *et al.*, 2002). This role seems to be locus specific since for the Igf2/H19 locus it was shown that nucleosome positioning did not depend on CTCF whereas CTCF dependent nucleosome positioning was observed in Kaposi's sarcoma associated herpes virus (KSHV) (Kanduri *et al.*, 2002, Kang *et al.*, 2013).

Binding sites for CTCF in mammalian DNA are abundant and binding sites of high occupancy are often conserved between tissues and species whereas low occupancy sites more likely to be tissue specific (Essien *et al.*, 2009). The human genome is suggested to contain between 15000 and 40000 CTCF binding sites based on computer predictions of which tens of thousands have been confirmed by ChIP experiments (Xie *et al.*, 2007, Kim *et al.*, 2007, Shen *et al.*, 2012, Barski *et al.*, 2007). The presence of CTCF binding sites on retrotransposons has been found to be an important factor in evolution (Schmidt *et al.*, 2012). Due to this feature the translocation of retrotransposons can form new boundaries between transcriptionally active and inactive chromatin which potentially lead to the activation or deactivation of whole gene clusters (Schmidt *et al.*, 2012).

CTCF has uncommon features like aberrant mobility on PAGE and the unusually high amount of 11 zinc fingers that enable it to bind large motifs (Filippova *et al.*, 1996). Even though the CTCF gene encodes for a product of 82 kDa, the CTCF protein migrates at 130 kDa or 70 kDa on a denaturing polyacrylamide gel. Further investigation revealed that the 130 kDa band seen on denaturing polyacrylamide gel was in fact the full length 82 kDa protein which inherently migrates aberrantly (Klenova *et al.*, 1997). The 70 kDa band is a C-terminally truncated version of the full length CTCF protein. This truncation was shown to

be mediated by the interaction of a sequence in the UTR with a sequence in the coding region of the CTCF mRNA (Klenova *et al.*, 1997). The 70 kDa isoform of CTCF is expressed in some cell lines and is able to interact with the 130 kDa form of CTCF, resulting in enhanced transactivation (Klenova *et al.*, 1997, Klenova *et al.*, 1993). However full length CTCF (130 kDa) is present in all cells and therefore the term CTCF, if not defined otherwise, refers to the full length variant of CTCF that migrates at 130 kDa. Furthermore dimerisation and oligomerisation of CTCF proteins has been observed *in vitro* and *in vivo* which have been suggested to be involved in DNA looping (Yusufzai *et al.*, 2004, Phillips and Corces, 2009, MacPherson and Sadowski, 2010).

Also a paralog of CTCF exists that shares the same DNA binding motif with CTCF but antagonises its function (Loukinov *et al.*, 2002). This paralog is called CCCTC-binding factor-like (CTCFL) or Brother of the Regulator of Imprinted Sites (BORIS) and was found to play a role in germ line cells and embryogenesis. Additionally low levels of BORIS expression were found in some tissues in adults, including the skin, testis and cervix (Rosa-Garrido *et al.*, 2012).

Full length CTCF consists of 3 major functional parts: an N-terminal region that can be post-translationally modified, the DNA binding middle region and the C-terminal region that can also be modified. Post-translational modifications include phosphorylation at the N-terminus and poly-(ADP)-ribosylation at the C-terminus. Phosphorylation of the N-terminus has been shown to switch the function of CTCF from a transcriptional repressor to a transcriptional activator (El-Kady and Klenova, 2005, Klenova *et al.*, 2001). Conversely a mutant in which phosphorylation was disturbed showed more efficient repression of transcription at CTCF binding sites (El-Kady and Klenova, 2005). Poly-(ADP)-ribosylation of CTCF has been shown to be essential for its function as an insulator. The lack of this modification caused abrogation of the barrier function at the chicken HS4 β-globin FII insulator element (Aker *et al.*, 2010). Additionally CTCF can be modified by the small ubiquitin like protein SUMO at either terminus (MacPherson *et al.*, 2009). This modification does not interfere with the DNA binding capability of CTCF but it enhances the repressor function of CTCF at the c-myc P2 promoter (MacPherson *et al.*, 2009).

The 11 zinc fingers of the DNA binding region recognize a primary and a secondary motif, both of which are present at a single CTCF binding site. In some locations the primary binding motif is able to define a CTCF binding site on its own but the presence of the secondary binding motif enhances binding and is sometimes required to facilitate CTCF

binding (Schmidt *et al.*, 2012). The primary motif is very large and consists of 20 bases. The secondary motif starts about 5 nucleotides downstream of the primary motif and consists of 9 bases. Also additional bases adjacent to the motifs are necessary for CTCF binding (Nakahashi *et al.*, 2013). Accordingly, CTCF binding sites are remarkably large and a DNAse footprint of CTCF was shown to cover about 50 nucleotides (Chau *et al.*, 2006). The 11 zinc fingers of the DNA binding region of CTCF do not all contribute equally to DNA binding as seen in Figure 10 (Nakahashi *et al.*, 2013). Zinc fingers 4-7 target the most important base pairs of the binding motif and bind to 80 % of CTCF binding sites. Zinc fingers 1-2 and 8-11 stabilise CTCF binding by targeting non-conserved flanking sequences if only the primary motif is present at the CTCF binding site in question. If a secondary motif is present zinc fingers 9-11 bind to this secondary motif (Nakahashi *et al.*, 2013). Conversely the existence of a tertiary motif in the vicinity of a CTCF binding site has been suggested that is able to destabilise CTCF binding (Nakahashi *et al.*, 2013). In this case zinc fingers 1-2 lose their affinity for DNA or get outcompeted by an unknown factor. In response to this, zinc finger 8 clusters with zinc fingers 9-11 at the secondary motif. If the secondary motif is absent, zinc finger 3 binds to the core motif and becomes essential for binding. CTCF has been shown to be an important protein in the control of DNA methylation and genomic imprinting since it

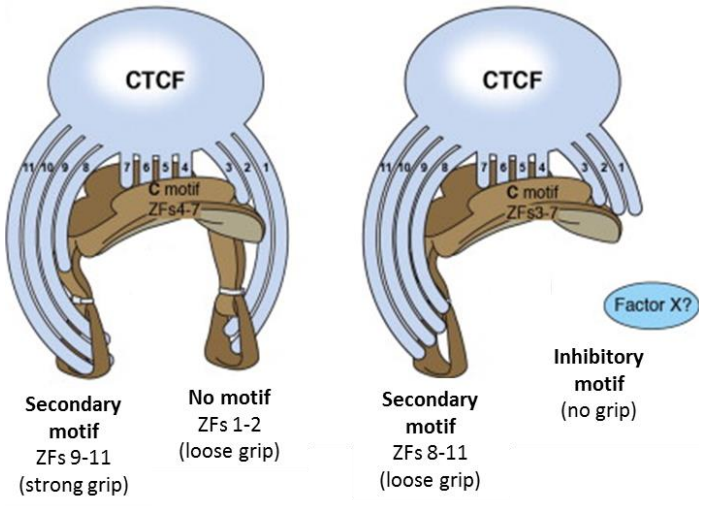


Figure 10) Saddle-model of CTCF multivalency
Left: Core motif with secondary motif. Strong binding of zinc fingers 4-7 to the core motif (seat of the saddle). The upstream motif is recognised by zinc fingers 9-11 and stabilises overall binding (strong grip). Zinc fingers 1-2 associate with DNA sequences that lack a consensus motif and stabilise binding to a lesser extend (loose grip).
Right: Core motif without secondary motif but with suggested factor X. Zinc fingers 1-2 lose affinity for DNA or get outcompeted by an unknown factor (X). This results in the clustering of zinc finger 8 with zinc fingers 9-11. If the secondary motif is absent, zinc finger 3 binds to the core motif and becomes essential for binding.
The Figure was modified from Nakahashi et al. with permission from the copyright holder (Nakahashi et al., 2013).

protects a region of at least 2 kbp in its vicinity from *de novo* methylation (Fedoriw *et al.*, 2004, reviewed by Klenova and Ohlsson, 2005). The methylation protection function of CTCF is facilitated via the protein poly (ADP-ribose) polymerase 1 (PARP1). This protein is able to regulate the activity of other proteins, like the methyltransferase DNMT1, by the addition of poly (ADP-ribose). CTCF forms a complex with PARP1 and the methylating enzyme DNMT1 which triggers PARP1-mediated poly-(ADP)-ribosylation of Conversely CTCF binding was shown to be prevented by methylation of the CTCF binding site (Singh *et al.*, 2012, Fedoriw *et al.*, 2004). However methylation based abrogation of CTCF binding may be locus specific since it was shown that CTCF can also bind to its binding site regardless of the fact that it is methylated or not (Salamon *et al.*, 2009). The methylation specific gene expression mediated by CTCF is especially important for imprinted loci and was first discovered at the mouse Igf2/H19 locus (Bell and Felsenfeld, 2000, Hark *et al.*, 2000, Kanduri *et al.*, 2000). Methylation on the paternal imprinting control region (ICR) blocks CTCF binding with the result of silencing the H19 gene while activating transcription from the Igf2 gene (Figure 11). On the maternal allele the ICR is not methylated so CTCF can bind and activate H19 while simultaneously repressing Igf2 (Banerjee *et al.*, 2001). This transcription regulation was shown to be a result of enhancer blocking. If CTCF is present in between an enhancer and a promoter it is able to block the

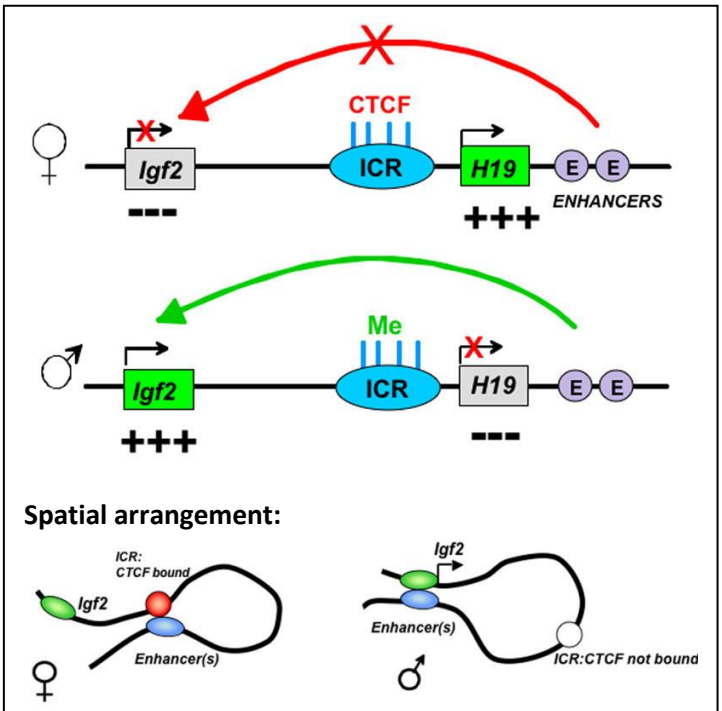


Figure 11) CTCF function in the Igf2/H19 locus
CTCF regulates gender specific transcription at the Igf2/H19 locus. CTCF binding at the internal control region (ICR) at the maternal locus blocks enhancers from the Igf2 promoter, resulting in H19 transcription. On the paternal locus the ICR is methylated so CTCF cannot bind. The enhancers are not blocked and act on the Igf2 promoter instead of the H19 promoter. CTCF-mediated looping is the basis of this mechanism. Figure from Wallace and Felsenfeld, reproduced with permission (Wallace and Felsenfeld, 2007).

enhancers from acting on the promoter (Bell *et al.*, 1999, Hark *et al.*, 2000, Recillas-Targa *et al.*, 2002). In this case the enhancers are free to target a secondary promoter. With the use of novel techniques for investigating the spatial arrangement of DNA it was found that CTCF is a major factor in the 3 dimensional structuring of DNA that was found to have an impact on gene regulation (Phillips and Corces, 2009). Forming 3-dimensional DNA conformations may be a result of the interaction of CTCF with structural nuclear proteins like nucleophosmin/B23 and the nuclear matrix (Dunn *et al.*, 2003, Torrano *et al.*, 2006). CTCF is able to bring two distant DNA loci in close proximity. This works in the form of a loop if the loci are on the same chromosome or in form of a bridge if the loci are on different chromosomes (Botta *et al.*, 2010). Chromatin interaction analysis with paired end tag sequencing (ChIA-PET) revealed 1500 intrachromosomal and 300 interchromosomal connections mediated by CTCF, all of which are non-random (Handoko *et al.*, 2011, Ling *et al.*, 2006). Furthermore, a genome wide DNA interaction map was created using the method Hi-C (Lieberman-Aiden *et al.*, 2009). Via intrachromosomal connections (loops) CTCF is able to separate active from inactive chromatin, to trap enhancers together with promoters in a loop or insulate genes by preventing a spatial connection of enhancer and promoter (Handoko *et al.*, 2011, Comet *et al.*, 2011).

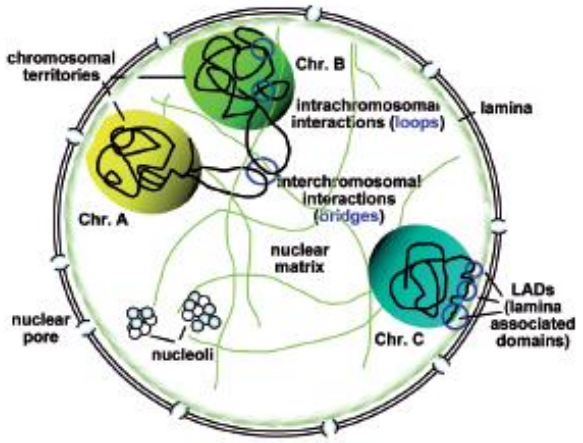


Figure 12) CTCF-mediated chromatin positioning in the nucleus
CTCF forms non-random interchromosomal connections of particular loci in the transcriptionally active centre of the nucleus. Additionally CTCF binds at the borders of transcriptionally silent lamina associated domains (LADs). Figure from Zlatanova and Caiafa, reproduced with permission (Zlatanova and Caiafa, 2009b).

The positioning and partitioning of DNA inside the nucleus is also suggested to be mediated by CTCF as illustrated in Figure 12. The nuclear lamina is situated at the inside of the nuclear envelope and DNA in this region has been found to be of little transcriptional activity and CTCF is often flanking lamina associated domains (LAD) (Guelen *et al.*, 2008). LAD change dynamically during differentiation which is potentially mediated by CTCF (Peric-Hupkes *et al.*, 2010). Furthermore it was shown that small and gene rich chromosomes tend

to be situated more centrally in the nucleus which might be important for papillomavirus genome maintenance (Bolzer *et al.*, 2005, reviewed by Cremer and Cremer, 2001).

The co-localisation of CTCF and cohesin is important for connecting distant DNA loci (Stedman *et al.*, 2008, Xiao *et al.*, 2011, Wendt *et al.*, 2008, Rubio *et al.*, 2008, Parelho *et al.*, 2008). Cohesin is a protein complex involved in sister chromatid cohesin during mitosis where it forms a ring structure around two DNA helices (Nasmyth and Haering, 2005, Wendt *et al.*, 2008). In post-mitotic cells cohesin still binds to DNA and half of its binding sites overlap with CTCF (Wendt *et al.*, 2008, Rubio *et al.*, 2008, Parelho *et al.*, 2008). CTCF-cohesin mediated DNA looping was shown to be essential for the transcriptional regulation of a variety of loci including the loci for Igf2, β-globin, T cell receptor α and interferon ɣ (Nativio *et al.*, 2009, Hadjur *et al.*, 2009, Chien *et al.*, 2011, Seitan *et al.*, 2011).

CTCF binding has been shown to be important in the regulation of RNA polymerase II progression during gene transcription. When the RNA polymerase II moves a long a template strand and encounters a CTCF-cohesin complex the transcription process is stalled until the CTCF-cohesin complex is dislocated. This stalling favours the assembly of a splicing complex resulting in the inclusion of exons with weak splice sites upstream of the CTCF binding site (Wada *et al.*, 2009, Shukla *et al.*, 2011).

The function of CTCF at particular binding site has been shown to depend on factors bound in the close vicinity of CTCF (reviewed by Wallace and Felsenfeld, 2007, reviewed by Zlatanova and Caiafa, 2009b). For example the estrogen receptor co-occupies some CTCF binding sites. For these loci it was shown that the binding of CTCF is needed to facilitate transcriptional activation by the estrogen receptor (Ross-Innes *et al.*, 2011). Other factors that can influence the function of CTCF when bound in its close vicinity include RNA polymerase I II and III, VEZF1, YY1, SMAD, FOXA1 and Oct4 (reviewed by Weth and Renkawitz, 2011).

In conclusion, it is challenging to determine what purpose CTCF serves at particular binding sites considering the vast variety of mechanisms in which CTCF is involved.

1.8 The role of CTCF in the life cycle of other DNA viruses

1.8.1 Epstein - Barr virus

Epstein-Barr virus (EBV) belongs to the herpesviridae which are large DNA viruses. It is maintained episomally and can establish various types of latency, each with different properties in terms of growth and transformation of the host cell.

Screening the EBV genome for CTCF binding sites resulted in the discovery of CTCF binding sites upstream of the promoters Cp, Qp, Wp, and the transcription start site of two latency specific small nuclear RNAs called EBER1 and EBER2 (Day *et al.*, 2007, Tempera *et al.*, 2010). At the binding site close to Cp, CTCF was shown to activate transcription by interacting with the viral protein EBNA1. Also the transcription start site Qp is regulated by EBNA1 and CTCF. At this locus, CTCF binds 40bp upstream of the EBNA1 binding site which is in turn located 10bp upstream of the transcription start site. Mutation of this CTCF binding site resulted in decreased maintenance of EBV genomes (Tempera *et al.*, 2010). The decline in genome maintenance was suggested to be due to decreased transcription from Qp and was accompanied by a much delayed increase in transcription from the promoters Cp and Fp. This increase occurred after 8 weeks of culturing and was accompanied by a further decrease in transcription from Qp. After 16 weeks there was no detectable transcriptional activity of the Qp promoter and the promoter sequence was highly methylated. Thus CTCF was shown to prevent epigenetic silencing of Qp and prevent promiscuous transcription from certain promoters (Tempera *et al.*, 2010).

Furthermore CTCF has been shown to be an important factor involved in the regulation of latency-specific transcripts in the latency types I and III (Chau *et al.*, 2006). Epstein-Barr nuclear antigen 2 (EBNA2) is one of the EBV latency proteins that are involved in transformation of the host cell (Zimber-Strobl *et al.*, 1994). Deletion of a CTCF binding site between the origin of replication and the Cp promoter resulted in elevated EBNA2 mRNA levels of 3.5 fold in the latency types I and III (Chau *et al.*, 2006). Conversely, overexpression of CTCF in latency type III cells decreased EBNA2 mRNA levels. In contrast to the Igf2/H19 locus, CTCF binding in this particular region of the EBV genome is not affected by CpG methylation of the CTCF binding site (Salamon *et al.*, 2009).

Interestingly, EBV episomes inside the nucleus have been found to pair up like sister chromatids (Nanbo *et al.*, 2007, Kanda *et al.*, 2007). This mechanism was shown to depend on cohesin which is known to interact with CTCF. Thus CTCF may be involved. During

mitosis EBV genomes were found to attach to host chromosomes, presumably to ensure equal segregation of genomes in daughter cells (Kanda *et al.*, 2007). EBNA1 was shown to be crucial for this interaction which, again, interacts with CTCF.

1.8.2 Herpes Simplex Virus

Like in EBV, the episome of herpes simplex virus type 1 (HSV1) is stably maintained in its host cell and can transition between a latent or lytic stage. CTCF has been shown to be involved in the regulation of those two stages. The HSV1 genome is subdivided into regions of predominantly hyperacetylated or hypoacetylated histone H3. During latency the region containing latency associated genes is hyperacetylated whereas the region containing lytic genes is hypoacetylated (Kubat *et al.*, 2004a, Kubat *et al.*, 2004b). CTCF is important for maintaining this pattern and clusters of CTCF binding sites for maintaining the histone acetylation pattern are located in the gene for the latency associated transcript (LAT) and some repeat regions of the HSV1 genome. Dissociation of CTCF from its binding site in LAT is accompanied by hyperacetylation of the lytic genes and reactivation of the virus from latency (Ertel *et al.*, 2012). Additionally the CTCF binding sites situated upstream of the promoters ICP0 and ICP4 promoters were shown to block the LAT enhancer from acting on its target (Amelio *et al.*, 2006, Ertel *et al.*, 2012).

1.8.3 Herpesvirus Saimiri (HVS)

Herpesvirus Saimiri (HVS) is a primate gamma-2 herpesvirus that has been shown to use CTCF for episome maintenance and gene regulation during latency (Zielke *et al.*, 2012). A CTCF binding map of the 155 kilobasepairs HVS episome was created and two particular binding sites within the promoter region of the latency-specific transcript orf73/LANA were mutated. The CTCF mutants showed a reduction in expression of the orf73/LANA transcript and impaired proliferation of the host cell (Zielke *et al.*, 2012). Additionally episome loss was observed in the mutants which could be a response to the reduction in oerf73/LANA transcript as LANA is important for viral maintenance in dividing T-cells. It was suggested that CTCF is vital for maintaining the epigenetic state of this virus and the loss of epigenetic maintenance may have caused the change in gene expression. The lytic cycle was not impaired in either of the CTCF mutants (Zielke *et al.*, 2012).

1.8.4 Kaposi's Sarcoma Associated Herpes Virus (KSHV)

Just like the previous viruses, KSHV is maintained as a stable episome and is able to physically attach to mitotic chromosomes to ensure equal segregation of viral episomes to the daughter cells (Kanda *et al.*, 2007).

The latent and lytic genes in Kaposi's sarcoma associated herpesvirus (KSHV) are in close proximity to one another which makes the tight regulation of chromatin modification and enhancer targeting necessary. Co-localisation of three CTCF-cohesin complexes in the latency control region of KSHV was observed during latent infection (Stedman *et al.*, 2008). The binding of CTCF in this particular region is involved in maintaining latency since the induction of the lytic cycle coincides with the loss of CTCF binding in this region (Stedman *et al.*, 2008). This is confirmed by the fact that deletion of a CTCF binding site in in the latency control region resulted in the re-activation of the lytic genes (Stedman *et al.*, 2008).

There are multiple putative causes for the changes in viral gene expression observed. For example, the altered ratios of various splice variants of mRNA transcripts in the CTCF mutants (Kang *et al.*, 2013). Next to this CTCF is necessary for the initiation of transcription by recruiting the RNA polymerase II to KSHV promoters in the latency control region (Kang *et al.*, 2013, Chernukhin *et al.*, 2007). However CTCF was also shown to be necessary for repression of the viral lytic genes since dissociation of CTCF resulted in the induction of the lytic genes (Chen *et al.*, 2012b). In this mechanism CTCF is suggested to position cohesin at gene promoters, resulting in inhibition of RNA pol II engagement (Chen *et al.*, 2012b).

Furthermore, nucleosome positioning in KSHV is determined by CTCF which is in contrast to the Igf2/H19 locus where nucleosome positioning has been shown to be due to underlying sequence rather than CTCF binding (Kang *et al.*, 2013, Kanduri *et al.*, 2002). Apart from nucleosome positioning the maintenance of histone modifications in the LCR could be attributed to CTCF (Kang *et al.*, 2013).

1.9 Hypothesis and objectives

The goal of this study was to characterise a potential CTCF binding pattern in papillomaviruses and to investigate the function of CTCF in the papillomavirus (PV) life cycle. Many of the various functions CTCF can provide have been shown to be used by various large episomally maintained DNA viruses (Kang *et al.*, 2013, Zielke *et al.*, 2012, Tempera *et al.*, 2010, Ertel *et al.*, 2012, Chen *et al.*, 2012b). Even though CTCF is often involved in the regulation of large scale genomic interactions there are still many functions of CTCF that have the potential to play a role in small DNA viruses like HPV. For example the regulation of transcription, epigenetics, methylation and the alteration of splicing mediated by the stalling of RNA pol II at CTCF binding sites. Thus I hypothesise that CTCF plays a role in the papillomavirus gene regulation.

To investigate this hypothesis, the first objective was to create binding maps for CTCF of various PV genomes. Computer predictions of CTCF binding sites in 8 different PV were made using 3 different prediction tools. The binding sites found were confirmed by electrophoretic mobility shift assays (EMSA) and binding maps of 7 out of the 8 PV tested were created.

The second objective was to create a CTCF mutant of a particular binding site and to create a cell line maintaining this mutant. The mutant was created by site directed mutagenesis and the abrogation of CTCF binding was confirmed by EMSA. Mutant and wild type HPV18 were transfected into human foreskin keratinocytes (HFK) to create a cell line model of HFK maintaining wild type or mutant HPV18.

The third objective was to elucidate CTCF function by comparing the life cycle of wild type and mutant HPV18. For this purpose the HFK cell lines were differentiated using methylcellulose and gene expression throughout the differentiation process was monitored using western blotting. Furthermore organotypic raft cultures were grown from the HFK cell lines and gene expression throughout the whole raft was analysed by immunohistochemistry.

As a fourth objective the capability of episomal maintenance in mutant and wild type HPV18 was investigated. To facilitate this, the copy number of episomes against the copy number of host DNA was compared using qPCR. DNA extracted from cell samples taken over a time course served as a template for this analysis.

2 Materials and Methods

2.1 Computer predictions of CTCF binding sites

The DNA sequences for each papillomavirus screened were obtained from the PubMed database and the accession number for each sequence is given in Table 3. Predictions with the tool provided by the CTCF binding site database (CTCFBSDB) were processed using the website at http://insulatordb.uthsc.edu/. If the tool identified overlapping motifs only the motif with the highest score was recorded in the tables of the results section. The screening results from version 1.0 and 2.0 of the CTCFBDB tool are limited because only the best match to each position weight matrix (PWM) in an input sequence was given. All lesser matches were neglected. Also, CTCFBSDB tool version 1.0 was unable to handle line breaks in an input sequences. Each line break was accounted as a spacer which led to flawed results in some occasions. This bug was fixed in a patch of version 2.0 but the problem of neglecting other matches than the best to each PWM remained. Screenings with *Storm* did not have this problem. The *Storm* screenings were done in collaboration with Dr. Jesse Ziebarth from the University of Tennessee Health Science Centre, USA. The PWM and tools used in the screening are shown in Table 4.

The Essex tool did no use any PWM. Instead it screened an input sequence for matches to EMSA confirmed CTCF binding sites from a database. This tool is available online at http://bsproteomics.essex.ac.uk:8080/bioinformatics/ctcfbind.htm. The locations of binding sites found on the antisense strand of the input sequence are given in 5' to 3' direction of the antisense strand. This can be confusing because one expects the locations of the binding sites to be given relative to the 5' to 3' direction of the sense strand. Thus the locations of binding sites found on the antisense strand are corrected in order to give their position relative to the sense strand.

Class	Papillomavirus type	Accession Number
Bovine PV	BPV1	AB626705.1
High risk HPV	HPV16	NC_001526.2
	HPV16 114/K	EU118173.1
	HPV18	AY262282.1
	HPV31	J04353.1
Low risk HPV	HPV6b	NC_001355.1
	HPV11	FR872717.1
Beta HPV	HPV8	M12737.1
	HPV38	U31787.1

Table 3) Accession numbers of FASTA sequences for the prediction of CTCF binding sites

PWM	Name in tables	Name in tools	Published by	Used by CTCFBSDB 1.0	Used by CTCFBSDB 2.0	Used by Storm
	1.PWM	AC EMBL_M1	Schmidt *et al.*, 2012	no	yes	yes
	2.PWM	AC EMBL_M2	Schmidt *et al.*, 2012	no	yes	yes
	3.PWM	AC REN_20	Kim *et al.*, 2007	yes	yes	yes
	4.PWM	AC MIT_LM2	Xie *et al.*, 2007	yes	yes	yes
	5.PWM	AC MIT_LM7	Xie *et al.*, 2007	yes	yes	yes
AGCACCACCTGGTGGTA	6.PWM	AC MIT_LM23	Xie *et al.*, 2007	yes	yes	yes

Table 4) PWM used to screen all PV sequences for CTCF binding sites
The figures of each PWM were reproduced with permission from the copyright holders (Schmidt *et al.*, 2012, Kim *et al.*, 2007, Xie *et al.*, 2007).

2.2 Generation of fluorescein amidite (FAM)-labelled DNA fragments for EMSA

The generation of FAM-labelled fragments containing the CTCF binding site of interest was done in a two-step PCR reaction. In step one a fragment of approximately 200 bp was amplified from template plasmids containing the PV genome of interested (Table 5). The forward primer for this reaction had the M13-overhang added to its 5' end (sequence: 5'**TGTAAAACGACGGCCAGT**3'). In the second step this fragment was used as a template in a PCR reaction together with a FAM-labelled primer that was complementary to the M13 overhang. In this second reaction the same reverse primer as in the first reaction was used. This resulted in FAM-labelled fragments ready for use in EMSA. All PCR reactions were done using Master Mix S from Peqlab. Labelled fragments generated this way were used immediately or stored at -20°C and protected from light. The PCR reactions were carried out using the thermocycler T Personal from Biometra. Reaction details and temperature cycles are given below. The details of the primers used to generate the control fragments can be found in Table 6. All other primers are listed in the appendix in Table 38 to 47.

First PCR reaction (fragment generation):

Reaction components:

20.5 µl H2O
2 µl FW primer (10 µM)
2 µl RV primer (10 µM)
0.5 µl template plasmid
25 µl Master Mix S from Peqlab

Cycles:

94 °C	5 min	
94 °C	30 s	30 cycles
59 °C	45 s	
72 °C	45 s	
72 °C	10 min	

Second PCR reaction (labelling):

Reaction components:

45.5 µl H2O
2 µl FAM primer (10 µM)
2 µl RV primer (10 µM)
0.5 µl template from first PCR
50 µl Master Mix S from Peqlab

Cycles:

94 °C	5 min	
94 °C	30 s	30 cycles
56 °C	45 s	
72 °C	45 s	
72 °C	10 min	

HPV Type	Vector	Cloning site	Obtained from
HPV16	pBR322	*Bam*HI	Prof. Ethel-Michelle de Villiers
HPV16 L1	pUF3	N/A	Prof. Martin Muller
HPV16 114/K	puC19	*Bam*HI	Prof. Ethel-Michelle de Villiers
HPV18	pGEM2	*Eco*RI	*Dr.* Frank Stubenrauch
HPV31	pBR322	*Bam*HI	Prof. Lou Laimins
HPV6b	pUC19	*Bam*HI/*Eco*RI	Prof. Ethel-Michelle de Villiers
HPV11	pBR322	*Bam*HI	Prof. Ethel-Michelle de Villiers
HPV38	pUC19	*Eco*RI	Prof. Massimo Tommasino
BPV1	pBR322	*Bam*HI	Prof. Elliot Androphy

Table 5) Template plasmids for the amplification of DNA fragments

Primer name	Template	Fragment start	Fragment end	Fragment length	Primer sequence (5'-3')	Tm
Neg. control FW	BPV1 AB626705.1	4910	5051	160	**TGTAAAACGACGGCCAGT** CCCGCAGTATTGCCTCTAAA	60.2
Neg. control RV	BPV1 AB626705.1	4910	5051	160	AGCACAGCTGGTTCTGCTTC	60.7
Pos. control FW	Human C-myc promoter	41894249	41894403	173	**TGTAAAACGACGGCCAGT** GGGATCGCGCTGAGTATAAA	58.7
Pos. control RV	Human C-myc promoter	41894249	41894403	173	GGATCTCCCTTCCCAGGAC	59.2
FAM-M13(-21)	All forward primers	N/A	N/A	N/A	FAM-TGTAAAACGACGGCCAGT	57.7

Table 6) Primers for generating EMSA control fragments
The M13-tag at the 5' position of the forward primers is shown in bold.

2.3 Generation of CTCF protein

The *in vitro* expression of CTCF for EMSA was carried out using 3 different kits from Promega according to the manufacturer's manual, the TnT® T7 Rabbit Reticulocyte lysate (RRL) System, the TnT® T7 Coupled Wheat Germ Extract (WGE) System and the TnT® SP6 High-Yield Wheat Germ Protein Expression System. These kits combine transcription of mRNA from template DNA with subsequent translation of this mRNA in a single reaction. Best results were achieved with the TnT® SP6 High-Yield Wheat Germ Protein Expression System. The T7 polymerase kits used a template of full length human CTCF protein with a 10xhistidine (HIS)-tag at the N-terminus which was inserted in the plasmid pCi. In this plasmid the transcription of CTCF mRNA was controlled by the T7 initiation site (Farrar *et al.*, 2010). The same CTCF sequence inserted into the plasmid pDrive served as a template when using the TnT® SP6 High-Yield Wheat Germ Protein Expression System. Both templates were kindly provided by Dr. Dawn Farrar from the University of Essex (Figure 13). The DNA sequence of each plasmid and the amino acid sequence of the translation product

can be found in the appendix (page 203). The reaction mix and the incubation conditions for each kit are given in Table 7. The presence of CTCF translated with all kits was confirmed using western blotting with the anti-CTCF antibody from Abcam (ab70303) at a dilution of 1:3333. As a negative control for EMSA, luciferase was translated and the presence of luciferase was confirmed using the luminometer FLUOstar OPTIMA (BMG Labtech).

	TnT® T7 Rabbit Reticulocyte lysate (RRL)	TnT® T7 Wheat Germ Extract (WGE)	TnT® SP6 Wheat Germ Extract (WGE)
Translation mix	25 µl RRL	25 µl WGE	30 µl WGE
Kit buffer	2 µl	2 µl	none
Template	1 µg pCi T7 construct	1 µg pCi T7 construct	5 µg pDrive SP6 construct
RNAsin	1 µl	1 µl	none
T7 polymerase	1 µl	1 µl	none
Amino acids minus leucine	0.5 µl	0.5 µl	none
Amino acids minus methionine	0.5 µl	0.5 µl	none
ddH_2O	filled up to 50 µl	filled up to 50 µl	filled up to 50 µl
Incubation temperature	30°C	30°C	25°C
Incubation time	90 minutes	90 minutes	120 minutes

Table 7) *In vitro* generation of CTCF protein using 3 different kits from Promega
Two different amino acid master mixes were used in a single reaction, each lacking a particular amino acid. This was done to have all amino acids available in the reaction mix. No radiolabelled amino acids were needed since the DNA fragments used in EMSA were already FAM-labelled. CTCF expression worked best with the SP6 WGE kit.

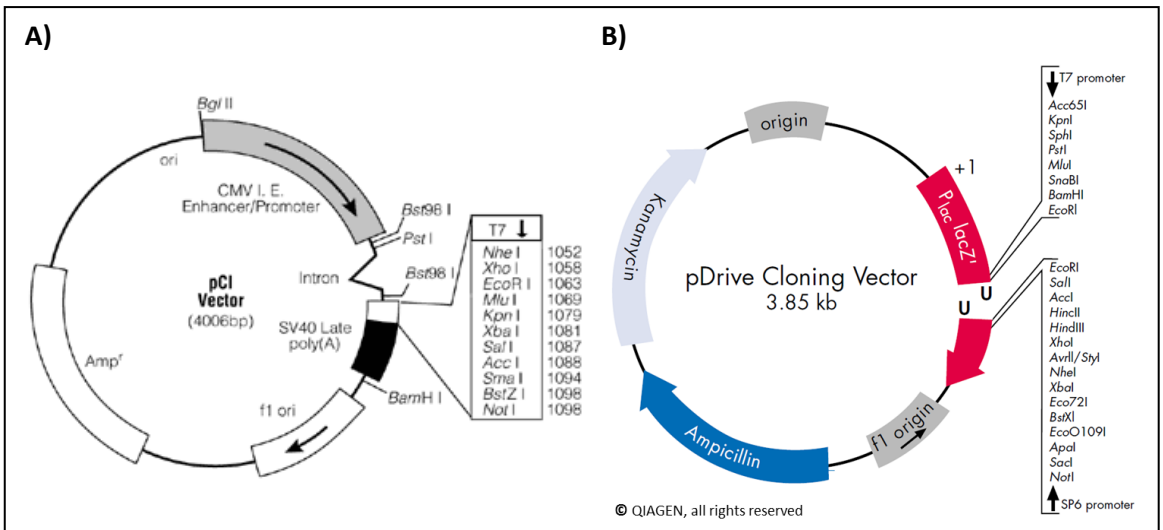


Figure 13) Plasmid maps of both vectors used for *in-vitro* expression of CTCF protein
These figures were reproduced with kind permission from Promega (**A**) and Qiagen (**B**).

2.4 Electrophoretic mobility shift assays

To facilitate binding of CTCF to its binding site, the previously generated FAM-labelled DNA fragments were incubated together with CTCF-containing WGE in a binding buffer. Two different binding buffers were used. Both giving the same results but one of them produced more defined bands. The first buffer was prepared as described in Sambrook *et al.* (Sambrook and Russell, 2006). The buffer that improved banding was a modified version of the buffer used by Zielke *et al.* (Zielke *et al.*, 2012). Both binding buffers consisted of a stock master mix and several other components that were added shortly before use. The composition of the stock master mixes are given in Table 8 and the composition of the final binding buffers are given in in Table 9.

20x Stock master mix defined by Sambrook *et al.*	Modified 10x stock master mix defined by Zielke *et al.*
1 M KCL	5 % Nonidet P40
400 mM HEPES-KOH pH 7.9	500 mM KCl
40 mM MgCl2	100 mM Tris-HCl pH 7.5
0.4 mM $Zn(O_2CCH_3)_2$ (zinc acetate)	
10 mM DTT	

Table 8) Stock master mixes used in EMSA
The master mix of Sambrook *et al.* was stored at 20°C whereas the modified 10x stock master mix from Zielke *et al.* was stored at room temperature

Complete binding buffer defined by Sambrook *et al.*	Complete modified binding buffer defined by *Zielke et al.*
1 µl of the 20x stock master mix (Sambrook and Russell, 2006)	1 µl of the modified 10x stock master mix (Zielke *et al.*, 2012)
0.4 µl of 1 µg/µl poly(deoxyinosinic-deoxycytidylic) acid sodium salt	1 µl of 1 µg/µl poly(deoxyinosinic-deoxycytidylic) acid sodium salt
4 µl of 50 % glycerol	1 µl of 50 % glycerol
2 µl of 20 % ficol 400	2.5ul of 20 % ficol 400
2 µl of labelled DNA fragment	2 µl of labelled DNA fragment
2 µl of CTCF containing T7 WGE	1 µl of CTCF containing SP6 WGE
8.6 µl ddH_2O	4 µl ddH_2O
	0.4 µl protease inhibitor cocktail
	0.1 µl of 100 mM PMSF
	0.01 µl of 100 mM DTT
Overall volume: 20 µl	Overall volume: 10 µl
Incubation time: 90 minutes	Incubation time: 60 minutes
Incubation temperature: 4°C	Incubation temperature: room temperature

Table 9) Reaction mixes and incubation conditions of both EMSA setups
The setup using modified buffer defined by Zielke *et al.* gave the highest quality data.

After incubation of the binding reactions the samples were loaded on 4.5 % native polyacrylamide gel. When preparing the gels it was made sure that all traces of ionic detergent were removed from the glass plates to avoid interference with the gel shift reaction. The gels were 1.5 mm thick and had 15 wells. Using a thick gel improved banding because the lack of a stacking gel caused the band to be as thick as the filling level of the well. A low filling level therefore resulted in sharper bands. When the gel had settled, the wells were labelled with a permanent marker. This was important since the samples did not contain any loading dye and it was therefore difficult to distinguish filled from empty wells. The gel casting and running equipment was the Mini-PROTEAN® Tetra Cell from BioRad with its accessories. The details of the gel composition as well as the running conditions can be found in Table 10.

Each fragment was incubated in 3 different conditions. The first condition included no WGE at all to ensure that the fragment showed a single band. The second condition included WGE in which luciferase was expressed. This ensured that no shift resulted from the components of the WGE. The third conditions contained WGE in which CTCF was expressed. Hence testing a single fragment occupied 3 wells of the gel. Each gel was loaded with two control fragments. The first was a positive control fragment of the C-myc promoter that was previously used in CTCF EMSA (Vetchinova *et al.*, 2006, Filippova *et al.*, 1996). The second was a fragment of the BPV1 genome containing no predicted CTCF binding site. Thus 6 out of the 15 wells of the gel were occupied with controls, leaving 9 wells which were sufficient to test 3 fragments. The gels were run in 0.5x TBE (1x TBE buffer: 89 mM Tris, 89 mM boric acid and 2 mM EDTA). The running conditions were different for the two experimental setups and can be found in Table 10. When the run was finished the gels were taken out of the glass plates and FAM fluorescence at 520 nm was detected directly from the gels using a Typhoon FLA 7000 (GE Healthcare Life Sciences).

Components and specifications	Volume and details
40 % polyacrylamide with a ratio of acrylamide to bisacrylamide of 29 to 1.	1.4 ml
5x Tris-borate-EDTA buffer (TBE)	1.25 ml
10 % ammoniumpersulfate (APS)	75 µl
TEMED	12.5 µl
ddH_2O	9.6 ml
Thickness	1.5 mm
Comb	15 wells
Running buffer	0.5x TBE
Running conditions for the buffer defined by Sambrook *et al.*	45 minutes at 150 V, on ice
Running conditions for the modified binding buffer defined by *Zielke et al.*	35 minutes at 150 V, room temperature

Table 10) Composition and specifications of native polyacrylamide gels used for EMSA
The volumes given are for a single BioRad mini-gel of 1.5mm thickness. The gels needed to be prepared fresh to ensure good EMSA quality.

2.5 Investigation of HPV18 mutations using a 60 bp probe

The FAM-labelled probes were delivered in the form of two complementary oligonucleotides that needed to be annealed before use. This was done by adding equimolar amounts of each oligonucleotide to a binding buffer consisting of 10 mM Tris-HCl pH8, 50mM NaCl and 1mM EDTA. This reaction mix was placed into a water bath of 98°C which was switched off and cooled down slowly overnight. The next morning the probes were ready for use. The EMSA using the probe was done as described the EMSA section with a few alterations. For the binding reaction 400 fmol of the probe was added to the binding reaction mix for each well. The native acrylamide gel was run at 150 V for 20 minutes. The positive control probe had been used in EMSA before (Bell *et al.*, 1999) and the sequences of both probes are shown below

Positive control probe:

5′-AGGCGCGCCCCCAGGGATGTAATTACGTCCCTCCCCCGCTAGGGGGCAGCGGCGCGCCT-3′

HPV18 binding site 2990 probe:

5′-GGCATACAGACATTAAACCACCAGGTGGTGCCAGCCTATAACATTTCAAAAAGTAAAGCA-3′

2.6 Bradford assay

Bradford assays were used to quantify the protein concentration of lysed human foreskin keratinocyte (HFK) samples in order to normalise the loading of polyacrylamide gels for western blotting. A 96 well plate setup was used and a standard curve was created using solutions of Bovine Serum Albumin (BSA) protein in concentrations of 0.9 mg/ml to 0.2 mg/ml in steps of 0.1 mg/ml. The Bradford reagent was prepared by mixing 1 part Bradford reagent with 4 parts ddH_2O. Each cell lysate was diluted 1:10. From the diluted lysates and the standard curve samples, 10 µl were pipetted into a well and 200 µl of Bradford reagent were added. All samples used were loaded in duplicate. Bradford reagent and samples were mixed by pipetting up and down. Subsequently this mixture was incubated at room temperature for 20 minutes and then analysed using a spectrophotometer at 595nm (BioRad). If a sample was outside the range of the standard curve the experiment was repeated using a more suitable dilution of this sample.

2.7 Western blotting

Western blotting was used to confirm the presence of *in vitro* generated CTCF and to compare gene expression in methylcellulose differentiated keratinocytes. Before the cell samples could be loaded on a gel they were lysed 300 µl of either SDS lysis buffer or urea lysis buffer (Table 11). The protein content of methylcellulose differentiated cells was determined by a Bradford assay according to the manufacturer's instructions. A normalised amount of cell lysate (40 µg) was mixed with loading buffer (Table 11) and heated to 99°C for 5 minutes. After the heat-denaturation the samples were loaded on a denaturing polyacrylamide gel. The samples of *in vitro* generated CTCF did not contain any cells and could therefore be directly mixed with SDS loading buffer, heated and loaded on the gel. The percentage of the gel used was depending on the protein to be analysed. The gel compositions can be found in Table 12. The gel apparatus used was the Mini-PROTEAN® Tetra Cell from BioRad and the gels were run for 25 to 60 minutes at 150 V depending on the percentage of the gel. The pre-stained protein markers used were Protein Marker IV and V from Peqlab. After running, the proteins in the gels were transferred to a PVDF membrane. This procedure was done using the Mini Trans-Blot® Module from BioRad at 400 mA for 90 minutes. The composition of the transfer buffer used is shown in Table 13. Afterwards the membrane was briefly washed in TBS/T before it was blocked with 5 % milk

dissolved in TBS/T for 45 minutes. Subsequently the appropriate amount of primary antibody (Table 14) in 10 ml of 5 % milk was applied to the membrane over a 1 hour incubation while shaking. The primary antibody was poured off and the membrane was washed in TBS/T 5 times for 7 minutes each. Afterwards the secondary antibody (in 10 ml of 5 % milk) was applied. Subsequently the blots were washed 5 times with TBS/T, each wash taking about 7 minutes. After washing the detection reagent SuperSignal West Dura Chemiluminescent Substrate (GE Healthcare Life Sciences) was used according to the manufacturer's instructions. Afterwards the membrane was analysed using a LAS 3000 (Fujifilm). Exposure time was chosen according to the strength of the signal and normalisation of protein content was confirmed by Ponceau staining after the transfer step and western blotting for Glyceraldehyde 3-phosphate dehydrogenase (GAPDH).

SDS lysis buffer	Urea lysis buffer	Loading buffer
1 % SDS	9 M Urea	32.9 mM Tris-HCl, pH 6.8
50 mM Tris-HCl, pH 7.4	50 mM Tris-HCl, pH 7.4	1.05 % SDS
5 mM DTT	5 mM DTT	13.15 % glycerol
1:50 protease inhibitor cocktail (PIC) added before use (Sigma-Aldrich, P8340)	1:50 protease inhibitor cocktail (PIC) added before use (Sigma-Aldrich, P8340)	0.005 % bromophenol blue

Table 11) Part 1 of buffers used for western blotting
The buffers are given in the concentrations at which they were used.

Components	Stacking gel	Resolving gel [8 %]
30 % polyacrylamide-bisacrylamide mixture in a ratio of 37.5 to 1.	500 µl	2.65 ml*
Tris-HCl	375 µl (1 M, pH 6.8)	2.5 ml (1.5 M, pH 8.8)
10 % SDS (sodium dodecylsulfate)	30 µl	100 µl
10 % APS (ammoniumpersulfate)	30 µl	100 µl
TEMED (N,N,N',N'-Tetramethylethylenediamine)	3 µl	10 µl
ddH_2O	2.05 ml	4.65 ml

Table 12) Composition of denaturing polyacrylamide gels
*The amount of ddH_2O and polyacrylamide-bisacrylamide mixture was adjusted to produce the desired percentage of the gel.

SDS running buffer	Transfer buffer	TBS/T
25 mM Tris-HCl, pH 8.3	25 mM Tris-HCl, pH 8.3	50 mM Tris-HCl, pH 7.6
192 mM glycine	192 mM glycine	150 mM NaCl
0.1 % SDS	5 % methanol	0.05 % Tween 20

Table 13) Part 2 of buffers used for western blotting
The buffers are given in the concentrations at which they were used.

Antibody	Supplier	Dilution	Secondary antibody	Protein MW [kDa]	Percentage acrylamide-bisacrylamide [%]
CTCF	Abcam (#ab70303)	1:1000	rabbit	82 (130)*	8
E2	Thierry Lab	1:1000	rabbit	42	10
E1^E4	r424**	1:3000	rabbit	10	15
E6	Santa Cruz (#sc-365089)	1:200	mouse	18	12
E7	Abcam (#ab100953)	1:100	mouse	12	15
Loricrin	Covance (#AF 62)	1:2000	rabbit	26	12
p53	D01**	1:100	mouse	53	10
GAPDH	Santa Cruz (#32233)	1:3000	mouse	37	12

Table 14) Antibodies used for western blotting
*CTCF is a protein of 82 kDa but migrates at 130 kDa. **Produced in the University of Birmingham.

2.8 Agarose gel electrophoresis

Agarose gel electrophoresis was used to analyse the DNA fragments used for EMSA, to confirm the template plasmids for fragment generation, to analyse the single steps of the site-directed mutagenesis procedure and to confirm DNA shearing of ChIP samples. Agarose was dissolved in TBE buffer (89 mM Tris-HCl, 89 mM boric acid and 2 mM EDTA. Final pH: 8.3) using a microwave. The percentage of an agarose gel was dependent on the size of the fragment to be analysed. Small fragments of 200bp required about 1.6 % agarose whereas plasmid-sized DNA was analysed on a 0.6 % gel. Ethidium bromide in the concentration of 0.8 µg per millilitre of gel was added to the solution after it had partly cooled down. Subsequently the gel was poured in a cast and left to set. Before loading the gel the DNA samples were mixed with 6x loading buffer containing 30 % glycerol and 0.25 % bromophenol blue. The gels were then inserted in an Agagel Mini Apparatus from Biometra and run at 120 V for 1 hour. DNA ladders used include Hyperladder II, IV and V (Bioline) as well as lambda *Hind*III ladder (Thermo Scientific).

2.9 Transformation and amplification of plasmids

Commercially available *Escherichia coli* (*E.coli)* JM109 were used for the transformation of plasmids. 33 µl of the chemically competent *E.coli* suspension was supplemented with 0.5 µl of plasmid solution (50-300 µg/ml) prior to incubating the mixture on ice for 20 minutes. Subsequently the samples were heated to 42 °C for 45 seconds and then chilled on ice for 5 minutes. To the chilled cells 100ul of Luria Bertani broth (LB) were added followed by incubation of the samples at 37 °C while shaking for 1 hour. The incubated samples were spread on agar plates containing the appropriate antibiotic for selection of the plasmid and incubated at 37 °C overnight. Single colonies were picked the next morning to inoculate 5 ml of LB medium to grow the culture overnight at 37 °C while shaking. Subsequently glycerol stocks were prepared by mixing 0.75 ml of the culture with 0.75 ml of 50 % glycerol solution before freezing the cells at -80 °C. The remaining culture was used to extract the plasmids using the QIAprep Spin Miniprep or Maxiprep kit from Qiagen according to the manufacturer's instructions. The plasmids purified this way were checked for sequence by restriction digests or direct sequencing. Throughout the transformation procedure a negative and a positive control was used. No plasmid was added to the negative control cells so these cells had no antibiotic resistance. The positive control cells were transformed with a known functional plasmid. Luria Bertani broth was composed of 10 g/L Tryptone, 5 g/L Yeast Extract and 5 g/L NaCl.

2.10 Chromatin immunoprecipitation

The ChIP-IT® Express Enzymatic kit from Active Motif was used according to the manufacturer's instructions with small deviations. In brief, un-differentiated HFK containing wild type and HPV18 genomes were grown to 70-80 % confluency, crosslinked for 3 minutes with 1 % formaldehyde diluted in culture medium, quenched with glycine for 5 minutes and harvested by scraping in 2ml PBS. The cells were then pelleted by gentle centrifugation (720 RCF at 4°C), supplemented with 1 µl PMSF (100 mM) and frozen at -80°C. The cells were thawed on ice and incubated in 1 ml lysis buffer for 45 minutes before douncing them on ice 270 times. The douncer used was the Kimble-Kontes part no. 885302-0002 with pestle B. The release of nuclei during the douncing process was monitored microscopically. After douncing the nuclei were pelleted (2,400 RCF, 10 minutes at 4°C) and resuspended in 350 µl digestion buffer. After 5 minutes of incubation at 37°C, 17 µl of

restriction enzyme solution (200 U/ml) were added to randomly cut the exposed DNA in-between nucleosomes. The digestion reaction was stopped after 10 minutes by adding 7 µl EDTA (0.5 M). This way the majority of the DNA fragments generated were between 150 and 450 bp in size. The debris in this chromatin sample was removed by centrifugation (18,000 RCF, 10 minutes at 4°C) and the chromatin-containing supernatant was used for ChIP. A sample of this chromatin served as input sample and for determination of the chromatin concentration.

To determine shearing efficiency and chromatin concentration the DNA had to be purified. The crosslinks between proteins were reversed by incubating 50 µl of the chromatin sample (supplemented with 150 µl ddH_2O and 10 µl 5 M NaCl) at 65°C overnight. RNA was removed from the sample using 1 µl RNase A (15 min at 37°C) and proteins were removed using 10 µl of Proteinase K solution (42°C for 1.5 hours). The DNA was purified using 200 µl phenol chloroform solution followed by centrifuging the sample for 5 minutes at 18,000 RCF (room temperature). The supernatant was taken off and supplemented with 20 µl sodium acetate (3 M, pH 5.2) and 500 µl ethanol (100 %) before incubating the sample at -80°C for at least 1 hour. Subsequently the sample was spun (4°C) at 18,000 RCF and the pellet was air dried before it was dissolved in 30ul ddH_2O. The DNA concentration was assessed using the Nanovue (GE Healthcare Life Sciences) and the shearing efficiency was assessed by agarose gel electrophoresis.

For the ChIP reaction, 15 µg of chromatin was supplemented with protease inhibitors (1 µl), protein G coated magnetic beads (25 µl), ChIP buffer 1 (10 µl), antibody suitable for ChIP (3 µl) and ddH_2O (filled up to 100 µl) followed by incubation overnight at 4°C on a rotator (Voss Model 4400). The beads were then washed once with 800 µl ChIP buffer 1 and twice with 800 µl ChIP buffer 2 followed by elution using 50 µl elusion buffer AM2. After 15 minutes of incubation on an end-to-end rotator, 50 µl reverse crosslinking buffer were added to the samples and the beads were removed. At this point the input samples (10 µl chromatin samples) were supplemented with 2ul NaCl and 88 µl ChIP buffer 2. ChIP samples and input samples were heated to 95°C for 15 minutes using the thermocycler T personal from Biometra. Subsequently 2 µl of proteinase K were added followed by incubation at 37°C for 1 hour to digest remaining proteins in the sample. Proteinase K was then deactivated using 2 µl Proteinase K stop solution and the samples were cleaned up using the GenElute™ PCR Clean-Up kit (Sigma-Aldrich). In the final step of the clean-up process, the ChIP samples were eluted with 300 µl of ddH_2O. At this stage the cleaned up

samples were ready for ChIP analysis. The polyclonal anti-CTCF antibody (pAb) was purchased from Active Motif, catalogue number 61311. The control antibody used was anti-FLAG antibody (FLAG M2 monoclonal, Sigma-Aldrich, F1804).

2.11 Analysis of CTCF binding using quantitative real time PCR

The purified ChIP samples were analysed using qPCR. The reaction mix for qPCR (Table 15) was prepared using the QIAgility pipetting robot (Qiagen). All reactions were set up in duplicate and every run was repeated at least three times. The primer pairs used are given in Table 16. The master mix used was the SsoFast EvaGreen supermix from BioRad which contained a polymerase that amplified DNA at the melting temperature of the primers. Thus no separate amplification step was needed in the temperature cycles. The qPCR analysis was done using the Rotor Gene (Qiagen) with the following temperature cycles:

qPCR reaction mix per well (10 µl)
5 µl SsoFast EvaGreen supermix (BioRad)
0.25 µl forward primer (10 pmol/µl)
0.25 µl reverse primer (10 pmol/µl)
2 µl ddH_2O
2 µl template

Table 15) qPCR reaction mix

- 5 minutes denaturing at 95°C
- 60 s anneal/extend 60°C
- 15 s melting 95°C

} 45 cycles

- melting curve: gradual temperature increase from 60 °C to 99 °C in steps of 1 °C

The melting curve of each primer pair was checked to ensure that a single target was amplified. The standard curves were generated in duplicates from 7 serial 1:5 dilutions of the wild type HPV18 input sample and their formulas were automatically calculated by the software of the qPCR machine. However the standard curves were checked manually for the hallmarks of a reliable qPCR which are a slope close to -3.33 and an R^2 value close to 1. Standard curves for each primer pair were imported into each experimental run with the same experimental setup. Each imported standard curve was adjusted to each particular run with the help of a control sample that was run with every qPCR experiment, including the original experiment to generate the standard curve. All qPCR experiments used the

same cycle threshold, so the Ct value of the control sample could be used to adjust the imported standard curve to each particular run.

Importing standard curves and primer efficiencies this way is common practice in many modern qPCR systems like the systems from Qiagen, Biorad and Applied Biosystems. This can be looked up in the manuals of these systems. The manual for the qPCR system used here can be found at http://www.qiagen.com/knowledge-and-support/resource-center/resource-download.aspx?id=bf47fa7f-6994-45af-b3bb-6e25b3ea7b17&lang=en.

The amount of template present in each ChIP sample could be quantified based on its Ct value. For normalisation the quantification of each fragment was expressed as a percentage of the input sample. A range of primer pairs across the whole genome was used to analyse the ChIP samples. If a local peak in the samples precipitated with anti-CTCF antibody appeared that was not seen in the FLAG samples a CTCF binding site was confirmed.

Fragment start	Fragment end	Fragment length	Primer name	Predicted binding sites on fragment	Tm [°C]	Sequence 5' to 3'
789	890	102	844 FW	5.PWM 844 score 9.98	59,5	AGTGTGAAGCCAGAATTGAGC
789	890	102	844 RV	5.PWM 844 score 9.98	60,0	ACCACGGACACACAAAGGA
1165	1256	92	1205 FW	E+1 1205	59,8	TTTAAAACGAAAGTTTGCAGGAG
1165	1256	92	1205 RV	E+1 1205	59,3	CTTGTAACCGTGGACTTAACTCTG
2167	2251	85	2200 FW	none	59,2	TTATAGGCGAGCCCAAAAAC
2167	2251	85	2200 RV	none	59,3	CCAATCTCCCCCTTCATCTAT
2968	3069	102	2990 FW	6.PWM 2990 score 23.4	59,4	GGGAACATGGCATACAGACA
2968	3069	102	2990 RV	6.PWM 2990 score 23.4	60,1	CTTGTAGGGCCATTTGCAGT
3519	3620	102	3550 FW	E-2 3488 and 6.PWM 3621 score 14.2	59,3	GGCACCGCAAAGACCTAC
3519	3620	102	3550 RV	E-2 3488 and 6.PWM 3621 score 14.2	60,2	ACCGAGAAGTGGGTTGACAG
4442	4536	95	4505 FW	1.PWM 4505 score 12.2 and 5.PWM 4538 score 10.5	59,7	GGTCGTACAGGGTACATTCCA
4442	4536	95	4505 RV	1.PWM 4505 score 12.2 and 5.PWM 4538 score 10.5	59,3	GGGCCCACAGGTTCAATA
5405	5506	102	5475 FW	E-1 5475	59,4	ACGGTCCCTTTAACCTCCTC
5405	5506	102	5475 RV	E-1 5475	59,2	GGCCGTGGGTGATACAAT
5712	5811	100	5768 FW	1.PWM 5768 score 12.9	59,0	TTTTATCATGCTGGCAGCTC
5712	5811	100	5768 RV	1.PWM 5768 score 12.9	59,2	CAGAAACCTTAGGAATATCCTGCT
6371	6462	92	6400 FW	none	59,6	GCAGCTTTTTGCTAGGCATT
6371	6462	92	6400 RV	none	59,8	CACGCATACCTGTGCCTTTA
7746	7845	100	7800 FW	none	59,1	ACTTTCATGTCCAACATTCTGTCT
7746	7845	100	7800 RV	none	59,5	ATGTGCTGCCCAACCTATTT

Table 16) Primers used for qPCR

The templates were ChIP samples containing the DNA of HPV18, accession number AY262282.1.

2.12 Quantification of viral episomes

Viral episomes were quantified using qPCR. The qPCR procedure is identical with the one described in the previous section. Modified versions of two mathematical methods were used for this quantification, the Pfaffl method and the Livak method (Pfaffl, 2001, Livak and Schmittgen, 2001). Those methods were originally designed for the quantification of cDNA from a test gene against a reference gene but they can also be used to determine the amount of viral genomes against the amount of a host allele. This is described in detailed in the results section and final formulas used are shown at the end of this section. Two primer pairs were needed, one targeting the viral episome and one targeting a single copy host locus, in this case the TLR2 locus (Gray *et al.*, 2010). The primers chosen are shown in Table 17. The templates used for the qPCR reaction with these primer pairs were extracted DNA from undifferentiated cells using the QIAamp DNA Mini Kit from Qiagen according to the manufacturer's instructions, including all optional steps.

Template	Primer name	Location	Fragment length	Sequence 5' to 3'	Tm
HPV18 AY262282.1	HPV18 Episome FW	2167-2251	85	TTATAGGCGAGCCCAAAAAC	59,2
HPV18 AY262282.1	HPV18 Episome RV	2167-2251	85	CCAATCTCCCCCTTCATCTAT	59,3
Human TLR2 locus	Human TLR2 FW	1893-1957	65	GCCAGCAAATTACCTGTGTGA	61.08
Human TLR2 locus	Human TLR2 RV	1893-1957	65	GGCGGACATCCTGAACCT	61.05

Table 17) Primers used for the determination of episome copy number
A fragment of the single copy locus TLR2 as well as a fragment of the HPV18 episome is targeted.

An important factor for calculating episome copy number is the primer efficiency of each primer pair used. The primer efficiency is the factor by which the amount of amplicon is increased with every cycle. Every sample was loaded in duplicate and every run was repeated at least three times. Standard curves for each primer pair were generated and imported as described in the previous section.

Formular for primer efficiency:

Primer efficiency (as factor) = $10^{\frac{-1}{slope}}$

In the Livak method the primer efficiency from both primer pairs used is assumed to be 2, resulting in a simplified calculation. The Pfaffl method takes individual primer efficiencies into account. The primer efficiencies from both primer pairs used were calculated from the standard curve and happened to be 2 for both primer pairs. In this case Livak method and Pfaffl method yield the same result. Thus the Livak method was used as shown below. The relative copy numbers of host allele and the viral episome were calculated from the Ct values of the qPCR samples. These Ct values were determined using the same threshold for each primer pair. Note that the copy numbers calculated this way were given relative to the number of host alleles.

Modified Livak method for the determination of the absolute episome copy number per host allele:

$$2^{-(Ct_{episome} - Ct_{host\ gene})}$$

Modified Livak method for the determination of the ratio of C3 genomes to WT genomes:

$$2^{-\left((Ct_{episome\ C3} - Ct_{host\ gene\ C3}) - (Ct_{episome\ WT} - Ct_{host\ gene\ WT})\right)}$$

2.13 Site directed mutagenesis of HPV18

The site directed mutagenesis was carried out using the QuikChange II XL Site-Directed Mutagenesis Kit from Agilent Technologies. Small deviations from the manufacturer's protocol were made. In brief, site directed mutagenesis primers were used that had three mismatches to the original sequence. These primers were very carefully designed to maintain a high melting temperature despite the mismatches while exhibiting a low degree of self-complementarity.

The sequence of the primer used for mutating the CTCF binding site at nucleotide 2990 was 5'-CAGACATTAAATCACCAGGTAGTGCCAGCCTATAACATCTCAAAAAGTAAAG-3' and the sequence of the primer used for mutating the CTCF binding site at nucleotide 5475 was 5'-CATTACCATCTACTACCGTTGTACGGCCCATTGTATCACCC-3'. The reverse primer for either reaction was the reverse complement of the mutagenesis primer.

The mutagenesis reaction was done on the vector pGEM2 containing the HPV18 genome. This was a PCR reaction performed according to the instruction manual using the DNA polymerase *PfuUltra* HF which is suitable for amplification of large targets. This way the vector including the HPV18 genome could be amplified as a whole. Remaining template was digested using the methylation-specific nuclease *Dpn*I, which digests the input template plasmid, but not plasmids that were amplified during the PCR reaction thus removing wild type plasmids from the PCR reaction. The mutated genomes were transformed into *E.coli* XL10-Gold ultracompetent cells and spread on an ampicillin-containing agar plate (100 µg/ml). After incubation overnight at 37 °C single colonies were picked to inoculate 5 ml LB broth supplemented with 100 µg/ml ampicillin. After another incubation overnight at 37 °C with shaking (200 rpm), the plasmids from these samples were extracted using the QIAprep Spin Miniprep Kit from Qiagen. These plasmids were sent for sequencing and plasmids containing the correct mutations were used for re-ligation of the viral genome.

2.14 Re-ligation of HPV18 genomes

Following site directed mutagenesis of the HPV18 genome cloned into pGEM2 vector, mutated plasmids were amplified in bacteria and purified using the QIAprep Spin Miniprep Kit (Qiagen). The purified plasmids were digested with *Eco*RI to separate the HPV18 DNA from the pGEM2 backbone. To facilitate this, the following components were combined in an Eppendorf tube: 10 µg of DNA of plasmid DNA, 2 µl of EcoRI (12 U/µl), 5 µl of Buffer H (Promega) and ddH_2O up to 50 µl total volume. This reaction mix was incubated at 37°C overnight. The following day, 1 µl of this reaction mix was separated on an agarose gel to ensure complete digestion of the plasmid. A band of 8000 bp for the linearised HPV18 genome was observed together with a band of approximately 2800 bp for the plasmid backbone pGEM2. The restriction enzyme was then heat inactivated by incubating the sample at 65°C for 20 minutes. Subsequently the sample was randomly ligated using T4 ligase. The following components made up the ligation mix: 650 µl ddH2O, 180 µl 5x ligation buffer (Invitrogen) and 10 µl T4 DNA ligase (3 U/µl, Promega, M1801). This ligation mix was incubated overnight in the cold room in a water bath set to 16°C. Subsequently the ligation products were precipitated. To do this the ligation samples were split into two Eppendorf tubes and the following reagents were added: 600 µl isopropyl alcohol and 180 µl of 5 M NaCl. The samples were then incubated at -80°C overnight. The next morning the ligation samples were spun down at 20,000 RCF for 30 minutes at 4°C and the supernatant was removed. The pellet was washed with 70 % ethanol followed by centrifugation at 20,000 RCF for 15 minutes. The supernatant was discarded and the pellet was air dried before suspending it in in 12 µl Tris-EDTA buffer (10 mM Tris-HCl pH 8 and 1 mM EDTA). The ligation product was analysed by separating 1 µl of the sample on an agarose gel. The presence of many random ligation products was confirmed, one of which was the re-ligated circular HPV18 genome. The ligation products were stored at -80°C.

2.15 Immunohistochemistry

Histological, paraffin-embedded slices of raft cultures were immersed in Histoclear for 10 minutes to remove the paraffin. This was followed by a bath in 100 % industrial methylated spirits (IMS) for 5 minutes, 3 washes in tap water and a bath in 0.3 % H_2O_2 for 15 minutes followed by another 3 washes with tap water. Subsequently the slides were immersed in 1 litre of antigen retrieval buffer (1 mM EDTA-NaOH pH 8.0 and 0.1 % Tween) and incubated overnight at 65 °C while stirring with a magnetic stirrer at 600 rpm. The next morning the slides were removed from the beaker and washed with tap water before blocking them with blocking buffer (20 % heat-inactivated goat serum (HINGS), 0.1 % BSA in PBS) for 1 hour at room temperature. Slides were then incubated with primary antibody (in blocking buffer) at 4 °C overnight (Table 18). The rafts were washed three times with PBS and incubated with secondary antibody (in blocking buffer) at 37 °C for 1 hour. Subsequently the slides were washed three times in PBS and incubated with 0.1 µg/ml Hoescht stain diluted in PBS for 10 minutes at room temperature. Following the incubation the slides were then mounted with three drops of Fluoroshield (Sigma) and a 40x20 mm coverslip and dried overnight at 4 °C. Imaging was done using a Nikon E600 microscope fitted with a 20x objective and pictures were captured using a Nikon DXM1200F digital camera.

Antibody	Supplier	Dilution factor	Secondary antibody	Colour
E1^E4	IDII (Roberts Lab)	1:5	mouse	red A594, green A488
E1^E4	r424 (Roberts Lab)	1:2000	rabbit	red A594
Cyclin B1	H-433 Santa Cruz	1:100	rabbit	red A594
BrdU	347580 Becton Dickinson	1:1000	mouse	green A488
Loricrin	PRB-145P Covance	1:1000	rabbit	red A594
Secondary AB green 488	A11029, Invitrogen, Alexa 488	1:1000	N/A	green A488
Secondary AB red 594	A11037, Invitrogen, Alexa 594	1:1000	N/A	red A594

Table 18) Antibodies used for immunohistochemistry

3 Results

3.1 Aim

This study aims to map CTCF binding sites to the papillomavirus genome, validate them and determine the function of CTCF for the papillomavirus life cycle.

3.2 Overview

CTCF binding sites in the BPV1 DNA sequence were predicted using 3 different *in silico* prediction tools. The binding sites found were experimentally confirmed by electrophoretic mobility shift assays (EMSA) using *in vitro* translated CTCF and fluorescein amidite (FAM) labelled DNA fragments of 200bp containing the CTCF binding site of interest. After confirmation or refute of predicted sites, additional fragments covering regulatory sequences such as promoters and the LCR were tested alongside DNA fragments of the BPV1 genome that were predicted to not bind CTCF. A CTCF binding site map for BPV1 was then created. Having mapped the sites in the BPV1 genome, predictions were made for 8 different HPV types. The binding sites found were tested by EMSA, giving rise to the creation of CTCF binding maps for all HPV types tested. Based on these binding maps a common binding pattern among HPV types was revealed. It was decided to mutate a particularly well conserved binding site in HPV18 at nucleotide 2990. This mutant was generated using site directed mutagenesis and the abrogation of CTCF binding was confirmed by EMSA. Mutant and wild type HPV18 were transfected into primary human foreskin keratinocytes (HFK) resulting in continuous cell lines. The binding of CTCF to the conserved CTCF binding site in these cell lines was confirmed using chromatin immunoprecipitation (ChIP) and the episome copy number over a time course was determined using quantitative real time PCR. Wild type and mutant HFK lines were differentiated in methylcellulose and viral gene expression upon differentiation was analysed by western blotting. Additionally Organotypic raft cultures grown from the HPV genome maintaining HFK lines were analysed and potential differences in morphology, gene expression and virus life cycle were monitored using immunohistochemistry and light microscopy.

3.3 Computer predictions of CTCF binding sites on 9 different PV types

3.3.1 Predictions using version 1.0 of the CTCFBSDB tool and the Essex tool

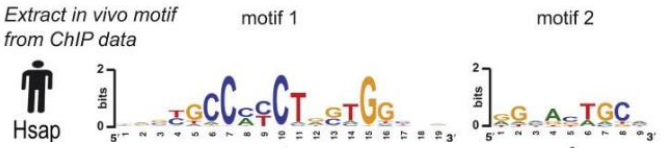
Predicting CTCF binding sites in various PV sequences was done using three different tools. The first tool was provided by the University of Essex and is based on EMSA-confirmed CTCF binding sites (link: http://bsproteomics.essex.ac.uk:8080/bioinformatics/ctcfbind.htm). It screens a target sequence against a database of known *in vitro* confirmed CTCF binding sites with a length of 10 nt. This programme works in two steps. The first step is to screen the sense strand of the input sequence for the motifs in the database. The second step is to search the antisense sequence of the input sequence for binding sites. The locations of the binding sites found in this second step are given in 5' to 3' direction of the antisense sense strand. This can be confusing due to the fact that all other predicted binding sites are given in 5' to 3' direction of the sense strand. Thus the results from this second step were corrected to unify the way of giving the location of predicted binding sites in 5' to 3' direction of the sense strand.

The second tool was provided by the CTCF binding site database (CTCFBSDB, version 1.0 – current version available at http://insulatordb.uthsc.edu/) and utilises 4 different position weight matrices (PWM) for CTCF binding site prediction (Xie *et al.*, 2007, Kim *et al.*, 2007, Bao *et al.*, 2008). During the course of this project there was a major update to the CTCFBSDB tool from version 1.0 to 2.0 which increased the number of PWM used to 6 (Figure 14) (Ziebarth *et al.*, 2013, Schmidt *et al.*, 2012). Shortly after the update an important bug fix was also released. However, the early predictions were carried out using version 1.0 of the CTCFBSDB tool and there was no update or bug fix applied at that time. The inherent flaws of version 1.0 of the CTCFBSDB prediction tool were not known at the time. Hence the first half of this chapter only covers predictions with the Essex tool and version 1.0 of the CTCFBSDB tool which are later complemented and corrected with CTCFBSTDB tool 2.0 and the third prediction tool, *Storm.*

The CTCFBSDB tool 1.0 screened for position weight matrices of 19-20bp in size (Kim *et al.*, 2007, Xie *et al.*, 2007). Every predicted binding site was rated with a score. A high score indicates a good fit to the PWM. Version 1.0 of the CTCFBSDB prediction tool contained a bug that made it count line breaks in an input sequence as spacers and therefore gave incorrect binding site locations and occasionally false predictions.

Despite not being aware of this bug at the time, it was realised that the location of a motif given by the programme did not exactly match up with the actual location of the motif. A correction step was therefore performed to give the correct location of each motif found. In Table 19 the location given by the tool is shown as the "native" position. The column labelled "corrected position" shows the real position of the motif recognised by the tool. During the course of this project it was also discovered that the CTCFBSDB tool had an even more serious flaw than the inability of handling line breaks. It gave incomplete results. However it is important to show the results given by the early, flawed version of the tool because these results led to constructive criticism which in turn led to a workaround of the tool.

1st and 2nd PWM, names in tool: **AC EMBL_M1 and M2**



3rd PWM, name in tool: **AC REN_20**

4th PWM, name in tool**: AC MIT_LM2**

LM2

5th PWM, name in tool: **AC MIT_LM7**

LM7

6th PWM, name in tool: **AC MIT_LM23**

LM23

AGCACCACCTGGTGGTA

Figure 14) Position weight matrices of CTCF
These PWM were used for predictions with the CTCFBSDB tool and *Storm.* Version 1.0 of the CTCFBSDB tool uses PWM 3-6 whereas version 2.0 and *Storm* use all PWM shown. There is no graphical representation of LM23 publicly available. The figures of each PWM were acquired with permission from the copyright holders (Schmidt et al., 2012, Kim et al., 2007, Xie et al., 2007).

Throughout this thesis, the naming of predicted binding sites found with either the CTCFBSDB or Essex tool follows a nomenclature system. The first character of the name always is a C or E, referring to the tool by which the binding site was predicted, so either C for the CTCFBSDB tool or E for the Essex tool, respectively. The second character is either a plus or a minus sign depending on whether the predicted binding site is on the sense (+) or antisense (-) strand. The third character is a number indicating the order of the binding site

on this particular strand. For example, the binding site E+2 is the second binding site on the sense strand that was found by the Essex tool. If binding sites are compared among different PV types, the virus type number is put in front of the name of the binding site, e.g. 18C-1 for the first binding site on the antisense strand found with the CTCFBSDB tool in the HPV18 sequence.

The results of predictions with the Essex tool and CTCFBFSDB 1.0 are summarised in Table 19 and Table 20. All 9 PV sequences screened are about 8kb in size and contain up to 12 predicted CTCF binding sites per sequence. The CTCFBSDB tool and the Essex tool rarely agree on particular binding sites, indicating the complexity and difficulty of predicting CTCF binding sites *in silico*. The CTCFBSDB tool frequently found overlapping motifs. If this was the case only the highest scoring motif is shown in the tables. Table 19 shows predicted binding sites in BPV1 before and after the correction process. Two motifs were found by the CTCFBSDB 1.0 tool compared to 9 motifs found by the Essex tool. All of these predicted binding sites were found in an area from nucleotide 2427 to 6682. This leaves a large area of about 3700bp including the whole LCR of BPV1 devoid of predicted motifs. Table 20 shows the results from screening 8 different HPV types with already corrected locations of the motifs. Almost all other PV types tested exhibited no predicted binding motifs in the LCR with the only exception of HPV38. Furthermore 3 out of 4 important high risk viruses harboured a very high scoring motif in the vicinity of nucleotide 3000. The low risk viruses HPV6b and HPV11 had a motif with similar score predicted in the vicinity of nucleotide 5400. Most of the other PV tested also had a predicted binding site in this area. However the binding sites in this area on PV other than HPV6b and HPV11 were mostly predicted by the Essex tool which does not give a score, making them difficult to compare without experimental data. Predicting CTCF binding sites in beta-HPV types revealed an accumulation of binding motifs in the vicinity of nucleotide 3600. Overall the Essex tool found substantially more motifs than the CTCFBSDB tool. The amount of motifs identified by both tools independently was low.

Name of binding site	Prediction tool	Orientation	"Native" start position	"Native" end position	Correct start position	Correct end position	Motif identified by tool	Reverse complement motif (only if BS was found on negative strand)	Score (only CTCFBSDB)
C-1	CTCFBSDB	antisense strand	3722	3741	3671	3690	TGAACCAGGTGGTGGTGCAG	CTGCACCACCACCTGGTTCA	10,29
C+1	CTCFBSDB	sense strand	6682	6700	6589	6607	TTAACAGTGGGGGACAATA	N/A	5,86
E+1	Essex	sense strand	3025	3034	3025	3034	CGCACAGAGG	N/A	N/A
E+2	Essex	sense strand	2427	2436	2427	2436	CAGGCAGAGG	N/A	N/A
E+3	Essex	sense strand	4535	4544	4534	4543	CGTCCAGGGG	N/A	N/A
E-1	Essex	antisense strand	7275	7284	663	672	CCGGGAGAGG	CCTCTCCCGG	N/A
E-2	Essex	antisense strand	5541	5550	2397	2406	CCAGGAGGGG	CCCCTCCTGG	N/A
E-3	Essex	antisense strand	6711	6720	1227	1236	CCTGGAGGGG	CCCCTCCAGG	N/A
E-4	Essex	antisense strand	2747	2756	5190	5199	TGTGTAGGGG	CCCCTACACA	N/A
E-5	Essex	antisense strand	3162	3171	4775	4784	TGCAGAGGGG	CCCCTCTGCA	N/A
E-6	Essex	antisense strand	3793	3802	4144	4153	TGAGCAGGGG	CCCCTGCTCA	N/A

Table 19) CTCF binding site predictions in the BPV1 reference sequence

The accession number of the BPV1 sequence screened is AB626705.1. Two binding sites were found using the CTCFBSDB tool 1.0. Native locations of the motifs found by the CTCFBSDB tool do not match up with the actual locations of the motifs in the sequence and therefore need to be corrected. The corrected locations are given in the columns next to the native locations. The Essex tool predicted 9 binding sites in this sequence and none of them matched up with the binding sites found with the other tool. Locations of binding sites found on the antisense strand with the Essex tool also needed to be corrected to unify the way of giving the location of a motif in 5' to 3' direction of the sense strand. The column next to the native locations shows the locations of the binding sites relative to the sense strand.

Class	PV	Corrected motif location	Prediction tool	Name	Motif identified by tool	Score (only CTCFBSDB)
High risk	HPV18	2990	CTCFBSDB	18C+1	AAACCACCAGGTGGTGCCAG	23,44
		1205	Essex	18E+1	CCATTAGGGG	N/A
		5474	Essex	18E-1	CATACAGAGG	N/A
		3488	Essex	18E-2	CGGTGAGGGG	N/A
	HPV16	2916	CTCFBSDB	16C+1	TAACCACCAAGTGGTGCCAA	20,05
		1283	CTCFBSDB	16C+2	AACTCAGCAGATGTTACAGG	10,18
		6512	CTCFBSDB	16C-1	GAACCACTAGGTGTAGGAA	8,61
		6127	Essex	16E+1	CCTATAGGGG	N/A
		5119	Essex	16E-1	CGCCTAGAGG	N/A
	HPV16 114/K	2917	CTCFBSDB	114kC+1	TAACCACCAGGTGGTGCCAA	27,63
		6128	Essex	114kE+1	CCTATAGGGG	N/A
		5120	Essex	114kE-1	CGCCTAGAGG	N/A
	HPV31	2413	CTCFBSDB	31C+1	AATGCACTAGATGGCAACC	14,68
		1278	CTCFBSDB	31C+2/E+3	AACGCAGCAGATGGTACAGG	13,98
		2332	CTCFBSDB	31C+3	CAACCACTGGCTGATGCTAA	8,34
		5179	Essex	31E+1	CCTTTAGGGG	N/A
		1093	Essex	31E+2	CATGCAGAGG	N/A
		1294	Essex	31C+2/E+3	CAGGTAGAGG	N/A
		885	Essex	31E+4	TGGGGAGGGG	N/A
Low risk	HPV6b	5425	CTCFBSDB	6bC-1/E-1	GCAGCCACAAGAGGGTGCAT	17,03
		5018	CTCFBSDB	6bC-2/E+1	CTATCACTAGATGATACCA	7,53
		4987	Essex	6bC-2/E+1	CCTATAGAGG	N/A
		1357	Essex	6bE+2	CATACAGAGG	N/A
		6099	Essex	6bE+3	TGGGCAGGGG	N/A
		5431	Essex	6bC-1/E-1	CCACAAGAGG	N/A
		6264	Essex	6bE-2	CCCAAAGGGG	N/A
		7206	Essex	6bE-3/E-4	CGAATAGAGG	N/A
		7257	Essex	6bE-3/E-4	CGTTTAGGGG	N/A
		4790	Essex	6bE-5	TGTGCAGGGG	N/A
	HPV11	5416	CTCFBSDB	11C-1/E-1	GCAGCCACTAGAGGGTGCAG	20,46
		6311	CTCFBSDB	11C-2	GTTCCAACGGGGGGCAGTC	13,21
		1357	Essex	11E+1	CATAGAGAGG	N/A
		5422	Essex	11C-1/E-1	CCACTAGAGG	N/A
		4781	Essex	11E-2	TGTGTAGGGG	N/A
		4058	Essex	11E-3	TGCAAAGGGG	N/A
Beta HPV	HPV38	3540	CTCFBSDB	38C+1/E+5	TCCCCCACCAGGGGAAGGAG	3,67
		3571	CTCFBSDB	38C+2	GGGACAGCAGAAGGCGGGG	5,77
		1331	CTCFBSDB	38C+3	CCGGCAGCAGAGGGTGATAT	15,57
		117	CTCFBSDB	38C-1	CTACCACAAGCGGTTGCTTG	10,87
		254	Essex	38E+1	CCTGTAGAGG	N/A
		3520	Essex	38E+2	CCTCCAGAGG	N/A
		4454	Essex	38E+3	CCTGGAGAGG	N/A
		3487	Essex	38E+4	CCAACAGGGG	N/A
		3544	Essex	38C+1/E+5	CCACCAGGGG	N/A
	HPV8	3642	CTCFBSDB	8C-1/E-2/E-4	TTCGCCTAGAGGTGGCTGTA	12.60
		2383	CTCFBSDB	8C-2/E-1	CAGCCAGCAGGGGTCTGTCA	12.65
		4037	CTCFBSDB	3C-3	CCATGACCACGTGGTGCTAA	6.50
		7361	Essex	8E+1	CCGTCAGGGG	N/A
		2390	Essex	8C-2/E-1	CCAGCAGGGG	N/A
		3650	Essex	8C-1/E-2/E-4	CGCCTAGAGG	N/A
		2346	Essex	8E-3	CACCAAGAGG	N/A
		3644	Essex	8C-1/E-2/E-4	CCTAGAGGTGGCTG	N/A

Table 20) Prediction of CTCF binding sites for various HPV types

The first three HPV types in this list have a binding site with a high score in the vicinity of nucleotide 3000. Both low risk HPV types tested harbour a binding site with a particularly high score around nucleotide 5400. The beta viruses have many CTCF binding sites predicted around nucleotide 3600. The accession numbers of sequences screened are: HPV18: NC_001357.1 (identical sites found in AY262282.1), HPV16: NC_001526.2, HPV16-114/k: EU118173.1, HPV31: J04353.1, HPV6b: NC_001355.1, HPV11: FR872717.1, HPV38: U31787.1 and HPV8: M12737.1.

3.3.2 The CTCFBSDB tool: Update, bug fix and remaining flaws

During the experimental confirmation of previously predicted binding motifs, a major update to the CTCFBSDB tool from version 1.0 to 2.0 was released. However the bug of including line breaks as spacers found in version 1.0 still persisted in version 2.0. Therefore, I reported the bug to Dr. Jesse Dylan Ziebarth from the University of Tennessee Health Science Centre, USA, who currently works on the tool (Ziebarth *et al.*, 2013). When the bug was discussed with Dr. Ziebarth, a more serious flaw of the tool became apparent. The tool only gave the best match of an input sequence to every single PWM and ignored all lesser matches. Hence it was giving incomplete results, especially when screening large sequences containing several CTCF binding motifs. This neglect of lesser binding sites had not been indicated anywhere on the CTCFBSDB website. Also a new flaw was introduced in version 2.0 of the CTCFBSDB tool through the implementation of two more PWM published by Schmidt *et al.* (Schmidt *et al.*, 2012). These PWM are used as if they independently determine CTCF binding motifs. However Schmidt *et al.* showed that those two PWM work together to define single CTCF binding sites as illustrated in Figure 15. Motifs matching the primary PWM can define a binding site on their own but the presence of the secondary binding motif a few nucleotides downstream of the first motif enhances the probability of CTCF binding. On the other hand the binding sites identified by the secondary PWM alone are meaningless without a binding site of the first PWM close by. This fact is not taken into account by the tool. Since the CTCFBSDB tool only showed the best match for each PWM it only gave one single match to the secondary PWM. Matches to this PWM are very common due to its small size of 9 nt which makes a single hit for this PWM not helpful. Considering the remaining flaws of the tool it was decided to predict CTCF binding sites with an independent method using the programme *Storm*.

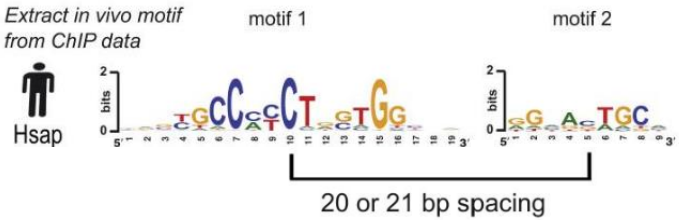


Figure 15) PWM of CTCF published by Schmidt *et al.*
This recently published position weight matrix consists of two separate motifs, both contributing to a single CTCF binding site. The distance between the mid-points of each motif is 20-21 base pairs in humans. Additionally Motif is able to define a CTCF site even in the absence of motif 1. Figure from Schmidt *et al.*, reproduced with permission (Schmidt et al., 2012).

3.3.3 Predictions using the programme *Storm* and 6 different PWM

Predicting CTCF binding sites with 6 different PWM on all 9 PV types using *Storm* was completed in collaboration with Dr. Jesse Ziebarth from Professor Yan Cue's research group at the University of Tennessee, Health Science Center, USA. Matches to the PWM are given a score but it is not known which score threshold indicates a true CTCF binding site. In this experiment we decided to choose the 10 best matches to each PWM for each individual PV type. This resulted in the identification of 540 potential CTCF binding sites. To reduce the work load a score threshold was needed to define the cut-off point from which a predicted binding site was tested for CTCF binding experimentally. It was decided to base the calculation for a threshold on the scores of the best 6 matches for each PWM from each of the 9 PV types, resulting in 54 different scores. The threshold was defined as the arithmetic mean of these 54 scores for every single PWM.

Table 21 shows the 6 best scores for each PWM and each virus and scores above threshold are highlighted in red. Predictions using the first and second PWM need to be addressed in particular. Those are the PWM published by Schmidt *et al (Schmidt et al., 2012)*. The 1st PWM downloaded (source: http://www.ebi.ac.uk/~schwalie/CTCFCell2012/) only covers the 14 bases whereas the PWM published in the paper consists of 19 bases. In a query with the authors it was stated that some of the nucleotides within the PWM described in Schmidt *et al.* did not have significant information content and were therefore left out of the PWM used for screening purposes. Hence the most important bases of the first PWM are covered by the 14 bases in the PWM used for screening, making this prediction valid. The secondary PWM published by Schmidt *et al.* was used as a whole. However due to its small size of 9 bases there were numerous hits across the whole input sequence and these hits are only significant if they are in close proximity to a motif identified by the first PWM. Having only the 10 best hits available made it difficult to set these motifs in relation to motifs matching the first PWM. It is likely that important hits were left out since they may not be among the 10 best matches. This fact, combined with the limited ability of the second PWM to predict CTCF binding sites by itself, led to the decision to exclude the matches to secondary PWM from further experiments. Based on the thresholds chosen it was decided to experimentally test the predicted CTCF binding sites shown in Table 22 by electrophoretic mobility shift assay (EMSA). Many of the motifs found by *Storm* were identified by multiple PWM. This reduced the overall number of fragments to be tested experimentally. Among many potential new binding sites found with *Storm*, a high scoring

binding site in the vicinity of nucleotide 3000 in HPV31 was found. Thus all high risk viruses tested had a conserved, high scoring CTCF binding motif predicted in this region. Many of the additional binding sites found with *Storm* were located in the L1 and L2 ORFs.

PWM	No.	HPV18	HPV16	HPV16 114/k	HPV31	HPV6b	HPV11	HPV38	HPV8	BPV1	Threshold (mean)
1.PWM (AC EMBL_M1)	1	18,41	13,90	19,27	19,16	16,92	18,44	15,27	14,29	9,35	10,72
	2	12,87	12,11	13,89	14,42	10,00	9,94	14,92	13,69	8,87	
	3	12,21	10,65	12,09	13,46	9,20	9,47	13,29	11,41	7,68	
	4	10,51	10,18	10,63	10,09	8,92	9,05	10,25	11,14	7,50	
	5	6,35	5,41	6,89	9,86	8,22	8,97	8,82	9,16	7,36	
	6	5,71	5,14	5,38	8,64	6,36	8,74	8,59	8,55	7,26	
2.PWM (AC EMBL_M2)	1	14,57	13,71	13,71	13,61	14,55	13,32	14,50	13,16	13,21	11,60
	2	14,57	13,42	13,42	13,01	11,67	13,32	11,46	10,33	12,39	
	3	12,19	11,31	11,31	11,26	11,37	11,50	11,03	10,31	12,29	
	4	12,19	11,20	11,20	10,48	11,09	10,70	10,59	10,17	12,14	
	5	11,90	11,11	11,10	8,97	10,68	10,61	10,02	9,74	11,48	
	6	11,90	10,69	10,67	7,87	10,34	10,30	9,85	8,49	10,40	
3.PWM (AC REN_20)	1	13,16	12,17	22,39	22,34	17,03	20,45	6,36	12,60	5,06	6,45 (mean of scores above 0 only)
	2	9,71	8,95	8,93	7,21	4,00	4,70	3,77	12,14	2,90	
	3	8,27	6,48	6,48	2,99	2,91	3,91	3,67	2,70	0,69	
	4	2,15	-0,30	-0,31	1,75	2,24	2,93	3,65	0,32	-1,50	
	5	0,56	-0,93	-0,94	-2,93	2,04	2,11	2,29	-0,33	-1,53	
	6	-0,14	-2,15	-2,17	-3,38	1,90	2,09	1,86	-1,64	-2,40	
4.PWM (AC MIT_LM2)	1	13,28	8,61	16,03	16,00	7,53	13,21	5,77	11,65	7,06	6,54
	2	7,27	7,68	8,60	14,68	6,89	11,50	4,07	6,55	6,46	
	3	7,26	6,18	7,67	9,01	3,78	7,33	4,03	6,34	5,86	
	4	4,97	5,44	6,18	8,23	3,55	7,29	3,79	4,82	5,47	
	5	4,90	4,22	5,71	6,38	2,29	4,85	3,44	4,22	1,70	
	6	4,22	4,21	5,43	6,25	2,16	4,77	3,10	3,91	1,24	
5.PWM (AC MIT_LM7)	1	16,65	10,18	18,21	18,93	12,90	18,61	15,56	12,65	10,28	8,71
	2	13,99	9,81	10,18	13,98	10,01	10,14	9,12	11,32	8,22	
	3	10,46	6,32	6,31	9,01	5,69	6,81	8,21	9,40	7,12	
	4	10,06	5,66	5,65	8,37	5,16	6,33	8,19	9,12	6,31	
	5	9,98	5,48	5,46	6,28	4,67	3,83	8,05	7,96	5,93	
	6	6,61	5,28	5,27	5,48	4,42	3,25	6,91	4,84	5,49	
6.PWM (AC MIT_LM23)	1	23,44	20,05	27,63	27,48	12,22	15,82	10,87	6,50	10,61	8,51
	2	14,25	9,32	9,31	8,34	8,06	9,30	9,15	6,35	8,52	
	3	12,19	9,08	9,07	7,48	6,80	5,42	8,49	5,58	7,99	
	4	7,71	5,96	8,20	7,01	6,50	5,16	5,38	5,55	7,58	
	5	7,19	4,89	5,93	5,56	4,73	4,76	5,07	4,61	5,62	
	6	7,15	3,66	5,64	4,53	4,59	3,35	4,75	4,54	4,69	

Table 21) *Storm*-scores of the 6 best matches of each PWM for each PV

The arithmetic mean of the scores for each PWM was determined and served as a threshold from which binding sites were experimentally tested. The second PWM (2.PWM) taken from Schmidt *et al.* has a large number of matches, all of which have a similar score. Matches to this PWM are only significant for CTCF binding if they are in close proximity to a match to the primary PWM defined by Schmidt *et al.* (1.PWM). However the secondary PWM was used independently of the primary PWM in this bioinformatics analysis. Hence it was decided to leave out results given by the secondary PWM for future experiments. The third PWM had negative scores among the 6 best matches. Those were considered not to be scores of real matches to the PWM and were left out of the calculation for the threshold. The name of each PWM as it was used in the prediction tools is given in brackets in the first column.

Class	Type	Start of predicted motif	PWM that identified this motif with a score above threshold [score]	Motif identified by highest scoring PWM
High risk	HPV18	844	5.PWM [10.0]	ATTCCAGCAGCTGTTTCTGA
		2990	1.PWM [18.4], 3.PWM [13.2], 4.PWM [13.3], 4.PWM [7.3], 5.PWM [16.6], 6.PWM [23.4]	AAACCACCAGGTGGTGCCAG
		3621	3.PWM [8.2], 5.PWM [14.0], 6.PWM [14.2]	TTGCCTGTAGGTGTAGCTGC
		4505	1.PWM [12.2]	CGTCCCCCAGTGGT
		4538	5.PWM [10.5]	GTAACAATAGATGGGTCTGT
		5768	1.PWM [12.9], 3.PWM [9.7], 4.PWM [7.3], 5.PWM [10.1], 6.PWM [12.2]	CACCACCTGCAGGA
	HPV16	1283	5.PWM [10.2]	AACTCAGCAGATGTTACAGG
		2916	3.PWM [12.2], 5.PWM [9.8], 6.PWM [20.1]	TAACCACCAAGTGGTGCCAA
		6515	1.PWM [13.9], 3.PWM [6.5], 4.PWM [8.6], 6.PWM [9.3]	CTACACCTAGTGGT
		6860	1.PWM [12.1], 3.PWM [9.0], 4.PWM [7.7], 6.PWM [9.1]	CTCCCCCAGGAGGC
	HPV16 114/K	1284	5.PWM [10.2]	AACTCAGCAGATGTTACAGG
		2917	1.PWM [19.3], 3.PWM [22.4], 4.PWM [16.0], 5.PWM [18.2], 6.PWM [27.6]	TAACCACCAGGTGGTGCCAA
		6516	1.PWM [13.9], 3.PWM [6.5], 4.PWM [8.6]	CTACACCTAGTGGT
		6861	1.PWM [12.1], 3.PWM [8.9], 4.PWM [7.7], 6.PWM [9.1]	CTCCCCCAGGAGGC
	HPV31	616	4.PWM [8.2]	ATAACAGTGGAGGTCAGTT
		1278	5.PWM [14.0]	AACGCAGCAGATGGTACAGG
		2413	1.PWM [14.4], 4.PWM [14.7]	AATGCACTAGATGGCAACC
		2854	1.PWM [19.2], 3.PWM [22.3], 4.PWM [16.0], 4.PWM [9.0], 5.PWM [18.9], 6.PWM [27.5]	TAACCACCAGGTGGTGCCAG
		6432	1.PWM [13.5], 3.PWM [7.2]	CTACACCTAGCGGC
Low risk	HPV6b	5018	4.PWM [7.5]	CTATCACTAGATGATACCA
		5425	1.PWM [16.9], 3.PWM [17.0], 4.PWM [6.9], 5.PWM [12.9], 6.PWM [12.2]	GCAGCCACAAGAGGGTGCAT
		6110	5.PWM [10.0]	CAGCCATTAGGTGTGGGTGT
	HPV11	4921	4.PWM [7.3]	CCACCTGTGGAGGCCAGTG
		5416	1.PWM [18.4], 3.PWM [20.5], 4.PWM [11.5], 5.PWM [18.6], 6.PWM [15.8]	GCAGCCACTAGAGGGTGCAG
		6311	4.PWM [13.2]	GTTCCAACGGGGGGCAGTC
		6636	4.PWM [7.3]	GAGCCACTAGGTGTATGTA
		6980	5.PWM [10.1], 6.PWM [9.3]	CCTCCACCAAATGGTACACT
Beta PV	HPV38	117	6.PWM [10.9]	CTACCACAAGCGGTTGCTTG
		690	6.PWM [9.2]	TTGCCACGAGGAGTTGCCTG
		1331	1.PWM [14.9], 5.PWM [15.6], 6.PWM [8.5]	CCGGCAGCAGAGGGTGATAT
		3543	1.PWM [13.3]	CTTCCCCTGGTGGG
		3573	1.PWM [15.3], 5.PWM [9.2]	CGCCTTCTGCTGTC
	HPV8	1959	5.PWM[9.1]	GCCCCAGAAGATGCTAATGC
		2387	1.PWM [13.7], 3.PWM [12.1], 4.PWM [11.7], 5.PWM [12.7]	AGACCCCTGCTGGC
		3642	3.PWM [12.6]	TTCGCCTAGAGGTGGCTGTA
		3769	1.PWM [14.3]	CTCCCCCTACTGCC
		5090	1.PWM [11.1], 4.PWM [6.5], 5.PWM [11.3]	GATACAGCAGGTGGCTGTTG
		5280	5.PWM [9.4	ACTACACCAGCTGGCTATGT
		6172	1.PWM [11.4	TGTCCCCTACTGAT
Bovine PV	BPV1	3670	5.PWM [10.3], 6.PWM [8.5]	GAACCAGGTGGTGGTGCAGT
		4258	4.PWM [7.1]	TGTCCACCAGATGTGATAC
		5317	6.PWM [10.6]	CTACCAACAGGTGTAGAAGA

Table 22) Storm-predicted CTCF binding sites with scores above threshold
Matches to 5 different PWM were taken into account and 9 PV were screened. All binding sites shown are analysed in subsequent experiments. Many motifs were recognised by multiple PWM.

3.4 EMSA validation of CTCF binding sites predicted with the CTCFBSDB tool 1.0 and the Essex tool

3.4.1 Testing predicted binding sites within BPV1 for CTCF binding

EMSA were performed to experimentally validate all predicted binding sites. In these assays a FAM-labelled DNA fragment (probe) of approximately 200 bp containing the binding site of interest was amplified by PCR and incubated in a binding buffer together with *in vitro* translated CTCF protein. This reaction mix was run on a native polyacrylamide gel and FAM fluorescence was detected at a wavelength 520 nm. On native polyacrylamide gels DNA fragments do not solely migrate according to their size, also the GC content plays a role. The higher the GC content the faster is the migration speed through the gel, so fragments of the same size can migrate differently. Free labelled DNA is seen as a single band on the gel. However if CTCF binds to the fragment, there will also be complexes of fragment and CTCF present in addition to the free probe. These complexes are considerably larger than the free probe resulting in a shifted, secondary band above the band of the free probe. This additional band is referred to as the band shift. The generation of a fragment containing the predicted CTCF binding site was done in two steps. The first step was a standard PCR-amplification of the fragment using a primer pair specific to the region of interest. Each forward primer used in the first step had an M13 5' overhang of 19 nucleotides to be used in the secondary PCR reaction (overhang: 5'-TGTAAAACGACGGCCAGT-3'). Fragments generated in the first PCR reaction acted as a template for a secondary PCR reaction. In this secondary reaction a FAM-labelled primer complementary to the overhang of the forward primer was used together with the same reverse primer used in the first PCR reaction. The sequence of the FAM primer was

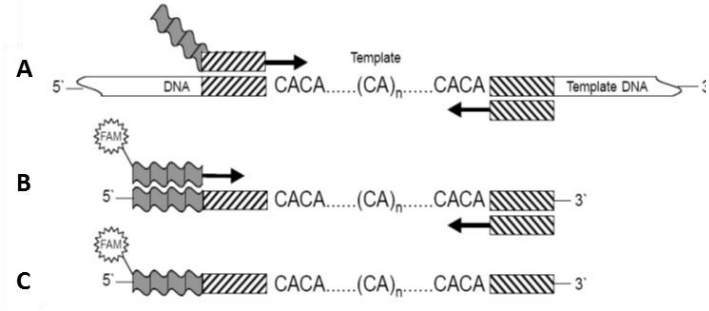


Figure 16) Generation of FAM-labelled DNA fragments using a labelled secondary primer
A) Amplification of DNA fragment from template DNA using an overhang primer. **B)** PCR with secondary, labelled primer targeting the overhang of the previous primer. **C)** Complete, labelled fragment ready to be used in gel shifts. Figure from Schuelke, reproduced with permission (Schuelke, 2000).

FAM-5'-TGTAAAACGACGGCCAGT-3'. As a result, all fragments generated this way were FAM-labelled using the same labelled primer in the secondary PCR reaction. The generation of labelled DNA fragments is illustrated in Figure 16.

Another crucial component of the binding reaction was CTCF protein. Two expression constructs for full length human CTCF with a 10x histidine tag (his-tag) at the N-terminus were kindly provided by Dr. Dawn Farrar from the University of Essex, UK (Farrar *et al.*, 2010). Both of these constructs contained the CTCF cDNA under control of either a T7 transcription initiation site or an SP6 promoter and could therefore act as a template for *in vitro* translation of CTCF using various kits from Promega that combine transcription and translation from template DNA in a single reaction. The kits used in this experiment were the TnT® T7 Coupled Reticulocyte Lysate System, the TnT® T7 Coupled Wheat Germ Extract System and the TnT® SP6 High-Yield Wheat Germ Protein Expression System.

Starting with CTCF translation using the reticulocyte lysate (RRL) kit, reticulocyte lysate-containing CTCF protein was combined with a labelled DNA fragment in the gel shift reaction. The presence of CTCF was confirmed by western blotting as seen in Figure 18. However it was discovered that the intensely coloured reticulocyte lysate interfered with the detection of FAM fluorescence and the band shift itself, resulting in a smear (Figure 17).

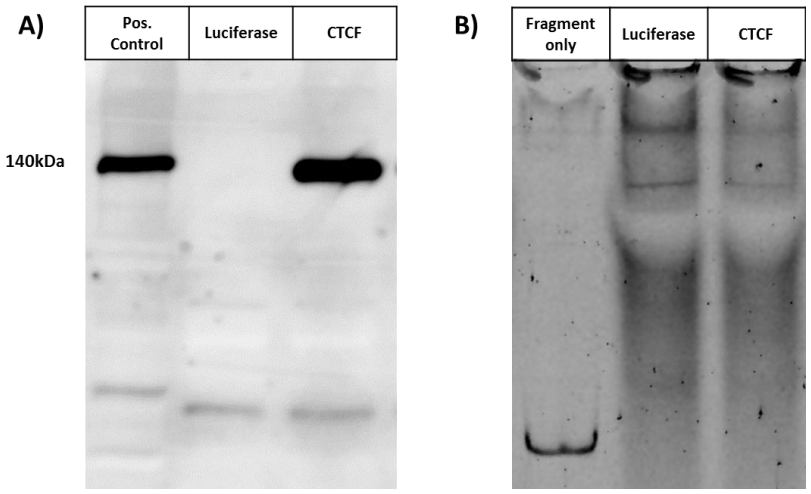


Figure 17) Western blot and EMSA of CTCF protein generated in RRL
Both images are representative of two independent experimental repeats.
A) Western blot for CTCF protein translated in RRL. The positive control is lysate of the mouse cell line ID13. The negative control is luciferase translated in RRL. The *in vitro* translated CTCF migrates slightly faster than the control CTCF. This may be due to the lack of post-translational modifications or a change in charge dues to the HIS tag. **B)** EMSA of the positive control fragment using CTCF and luciferase translated in RRL. Whole RRL containing luciferase or CTCF was incubated with the FAM labelled DNA and the complexes formed were separated by PAGE. Both lanes containing RRL with either protein show a smear whereas the lane containing fragment only shows a single band.

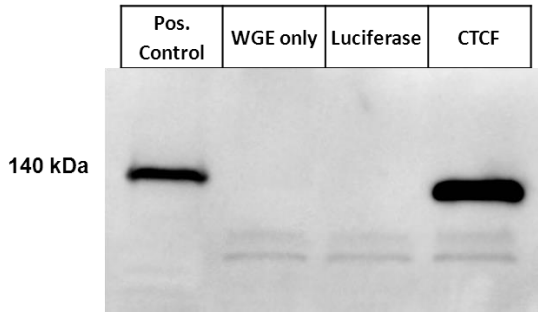


Figure 18) Western blot of CTCF protein generated in WGE
This image is representative of two independent experimental repeats. The positive control is cell lysate of mouse fibroblasts of the cell line ID13. The CTCF band from wheat germ extract migrated faster than the band from the positive control. This may be due to the lack of posttranslational modifications.

Using purified CTCF protein would have solved this problem. However, attempts to purify CTCF protein using a nickel column failed, most likely due to the 11 zinc fingers in CTCF protein that can bind to nickel, preventing efficient elution of the protein off nickel affinity resin used to purify histidine tagged proteins. Hence it was decided to *in vitro* translate CTCF in wheat germ extract (WGE) instead of reticulocyte lysate using another kit from Promega; the TnT® T7 Coupled Wheat Germ Extract System. WGE does not interfere with the detection of FAM fluorescence and CTCF-containing WGE could therefore be applied directly to the gel shift reactions without prior purification of CTCF protein. Successful expression of CTCF WGE was confirmed by western blotting (Figure 18). Whole cell lysate of the mouse fibroblast cell line ID13 was used as a positive control for CTCF western blots since this cell line was previously shown to express high levels of CTCF that cross-react with the human CTCF antibody used. The CTCF protein in WGE and RRL runs as a slightly higher band on Western blot than CTCF protein from ID13 cells. Nevertheless the DNA binding capacity was still confirmed as seen in the gel shifts (Figure 20). The addition of a histidine tag should rather decrease the migration speed instead of increasing it due to the added molecular mass. However the CTCF protein from the ID13 cell line migrates closer to 140 kDa than the expected 130 kDa. This may be due to the addition of potential posttranslational modifications. The SDS in the resolving gel should equalise the mass-to-charge ratio of the protein, but in the case of CTCF it has been shown that its charge can still influence its migration on a denaturing gel (Klenova et al., 1997). A change in charge due to the histidine tag should not have occurred since histidine (pKa 6) is de-protonated at the pH of the resolving gel (pH 8.8) and therefore not charged.

During a set of gel shift experiments, a drop in sensitivity occurred. This led to a large series of trouble shooting experiments targeting the composition of the binding buffer, its single components as well as every component of the *in vitro* expression kit. Low CTCF-protein

expression in T7 wheat germ extract kit was found to be the cause of the decline in EMSA sensitivity. Since all components of the WGE T7 kit were already replaced without any improvement of CTCF expression it was decided to use a different expression system based on the SP6 promoter; the TnT® SP6 High-Yield Wheat Germ Protein Expression System. Using protein generated with the SP6 system in combination with the best binding buffer found in the troubleshooting process for the T7 WGE generated CTCF protein, the overall quality of EMSA was increased. Due to the switch to another experimental setup, there are two separate methods with which the EMSA results were achieved. However, many of the fragments were tested with both methods and were found to give the same result.

The first method utilised 2 µl CTCF protein expressed in T7 WGE incubated with 2 µl of FAM-labelled DNA fragment in the EMSA buffer defined by Sambrook *et al.* given in Table 9 (materials and methods) (Sambrook and Russell, 2006). Incubation of WGE with the labelled DNA fragment was carried out at 4 °C for 1 hour followed by separation of the samples on a 4.5 % native polyacrylamide gel on ice. In the second method, CTCF protein was expressed in SP6 high yield WGE. From this 1 µl of WGE was incubated with 2 µl of FAM-labelled DNA fragment in the modified version of the buffer described by Zielke *et al.* shown in Table 9 (materials and methods) (Zielke *et al.*, 2012). The incubation was carried out at room temperature for 1 hour followed by separation of the samples on a 4.5 % native polyacrylamide gel run at room temperature. Both buffers contained Poly (deoxyinosinic-deoxycytidylic) acid sodium salt, an unlabelled nucleotide polymer that was added in excess to the DNA fragment in order to reduce non-specific binding of proteins to DNA. Experimental controls used in all EMSA included a fragment of the c-Myc promoter which is known to bind CTCF *in vivo* and had been used in EMSA before (Vetchinova *et al.*, 2006, Filippova *et al.*, 1996). The CTCF binding site of the positive control fragment is shown in Figure 19. The negative control was a fragment of BPV1 that contained no predicted CTCF binding sites at all.

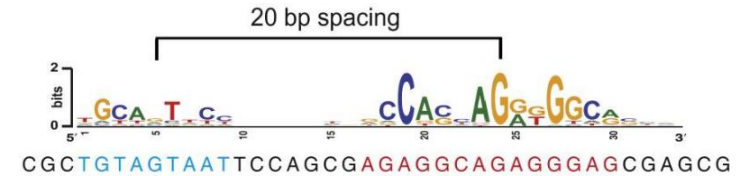


Figure 19) The CTCF binding motif in the human C-myc promoter
The positive control fragments used in EMSA contains this motif Figure from Schmidt *et al.*, reproduced with permission (Schmidt et al., 2012).

Since CTCF-containing WGE was used in gel shift assays, the possibility that other components of the WGE could cause a band shift needed to be addressed. To do this, luciferase protein was expressed in WGE and incubated with the fragment of interest as a secondary control. Thus it was possible to distinguish between background binding from WGE and binding that was specific to CTCF protein.

Some publications suggest pre-running gels before use to remove remaining ammonium persulfate and EDTA as well as bringing the gel to a constant temperature and equilibrating possible binding factors in the running buffer (Kothinti *et al.*, 2011, Mills and Marletta, 2005). However for both experimental setups used here pre-running decreased the overall EMSA quality rather than increasing it. Hence it was decided to use gels that were not pre-run.

Since not all predictions were carried out at the same time, the confirmation experiments for each prediction were also not all performed at the same time either. Instead, they were split into sections which are presented in chronological order to follow the train of thought and to justify why certain experiments were done. Starting with BPV1, fragments containing motifs predicted by either the Essex tool or the CTCFBSDB tool 1.0 were amplified from the BPV1 genome and labelled in a secondary PCR reaction. All these fragments were used for EMSA analysis and, at the very least, each experiment was carried out in duplicate. The summarised results are given in Table 23 of the following section. Figure 20 shows an example of an EMSA. In addition to the positive and negative control, 3 fragments each of which contains a predicted binding motif from the Essex tool are shown. Fragment Essex-3 is the only fragment that shows two weak band shifts so it is suggested that either two CTCF proteins can bind this fragment simultaneously or that a single CTCF protein can bind the fragment at two different and mutually exclusive locations leading to a distinct spatial conformation of the complex and therefore a distinct band shift for each possible binding location. Essex-4 shows no band shift. Fragment Essex-5 shows a shift that is stronger than the shifts of fragment Essex -3 but not as strong as the shift seen in the positive control. Therefore this shift is classified as medium strength. In some experiments a second band was seen in the lane containing free DNA only. Several fragments showed this secondary band occasionally but not consistently and the band always disappeared as soon as any WGE was added to the fragment no matter if the WGE contained luciferase or CTCF (Figure 21). When the same fragment that gave a secondary band on the gel shift was run on an agarose gel and stained with ethidium bromide, only a single band was seen.

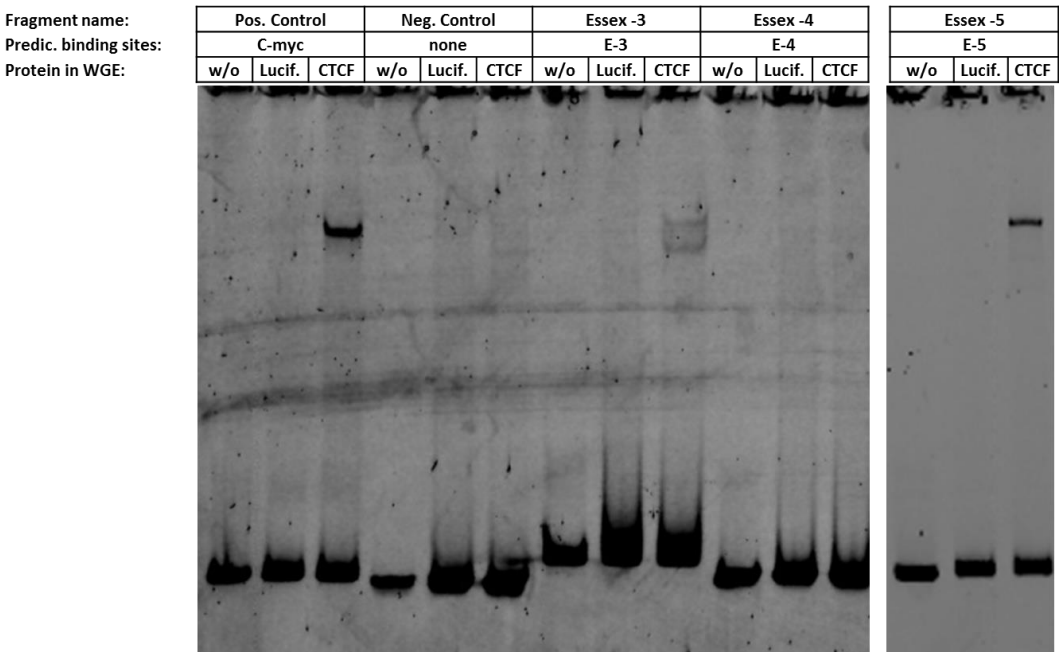


Figure 20) Example EMSA for testing predicted CTCF binding sites in BPV1
This image is representative of two independent experimental repeats. Fragment Essex-3 shows two weak band shifts whereas fragment Essex -5 shows a single but stronger shift. All remaining gels for testing BPV1 binding sites can be found in the appendix.

Hence it is suggested that the additional band was a complex of DNA fragments, most likely a homodimer consisting of two labelled DNA fragments. This complex may have formed in the reaction buffer but was disrupted when incubated with WGE.

As shown in Figure 21, fragment Essex+2 contains two predicted binding sites, which are in close proximity; E+2 and E-2. However, only a single weak band is seen on the EMSA with a DNA fragment that contains both of these predicted binding sites. From this data it was

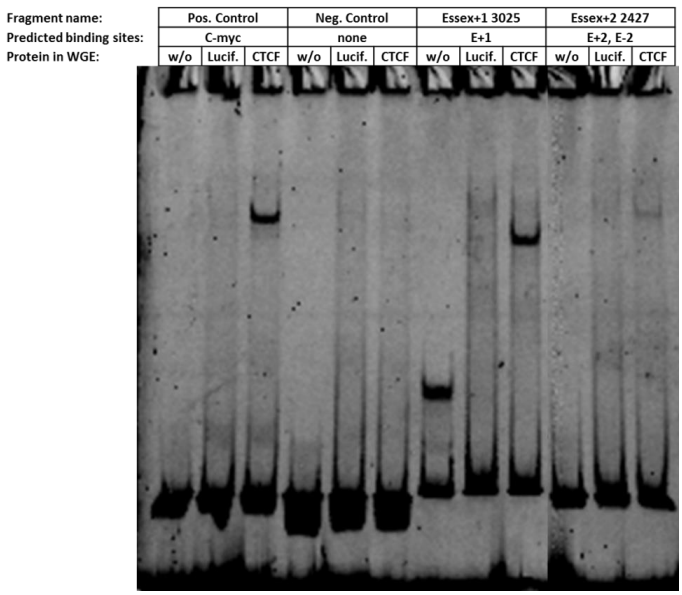


Figure 21) Typical nonspecific band in EMSA
A few samples showed a non-specific bands as seen in the well with Essex+1 DNA fragment only. This band disappeared upon addition of any WGE. A strong shift is seen with the Essex +1 fragment. The Essex +2 fragment shows a very weak shift.

not possible to distinguish CTCF binding to the E+2 motif or the E-2 motif. Fragment Essex+1 shows a single, strong band. Interestingly this fragment covers nucleotide 3000 which is the area where all of the high risk viruses have a predicted CTCF binding site with a high score. Among all the other BPV1 fragments with predicted binding motifs only C-1 showed a weak shift. The remaining fragments were all tested negative for CTCF binding. A complete collection of all gel shifts in duplicate can be found in the appendix.

3.4.2 Testing CTCF binding to additional BPV1 fragments covering the LCR and viral promoters

The findings from the previous EMSA revealed that many of the predicted CTCF binding sites were false positives. However, EMSA experiments may not always correctly resemble *in vivo* conditions, considering that these experiments are done on purely un-methylated DNA and that no other cellular factors are present. Nonetheless, they are good indicators for putative *in vivo* binding sites. As expected, predicting CTCF binding sites worked to a certain degree but it had its limitations. Hence it was decided to look for unpredicted CTCF binding sites in functionally important regions of the BPV1 genome. The whole LCR of BPV1 was screened for CTCF binding using overlapping fragments of 200bp. Each fragment was overlapping with the next by at least 30bp. A weak band shift was seen with fragment 6 of the LCR even though no binding site was predicted here (Figure 22).

BPV1 has three promoters outside the LCR of which two already have a gel shift confirmed binding site in their close vicinity. The only promoter outside the LCR that had not been tested for CTCF binding due to the lack of a predicted binding site is P890. Thus a fragment containing P890 was tested for CTCF binding resulting in a shift of weak strength as seen in Figure 22. Therefore every fragment containing a promoter sequence from outside the LCR was shown to have a band shift of weak strength. Also 7 random fragments and a fragment covering an E2 binding site outside the LCR were tested for CTCF binding. The fragment containing the E2 binding site does not show a shift. However one out of the 7 random fragments resulted in a weak shift as seen in Figure 22. Table 23 summarises all gel shift data of BPV1 up to this section. The EMSA from which this table is generated can be found in the appendix.

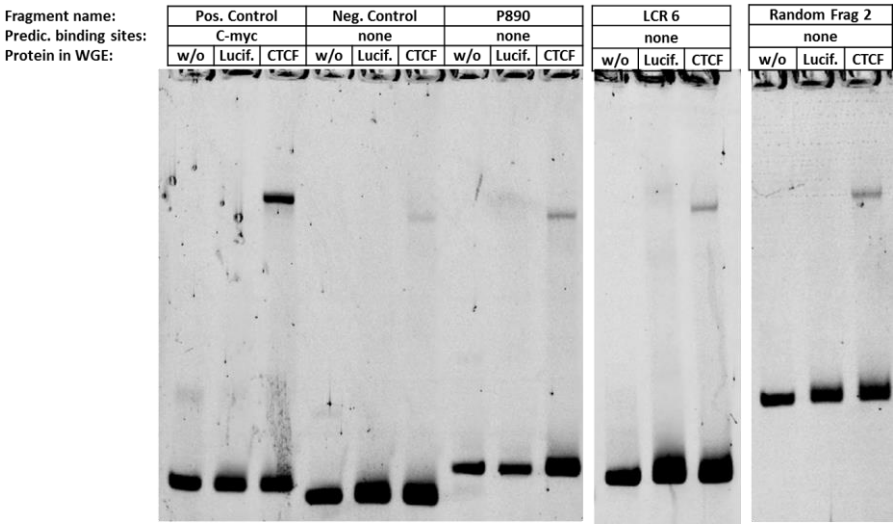


Figure 22) EMSA of functional elements and a random fragment
This image is representative of two independent repeats. The fragment containing promoter P890 shows a gel shift of weak strength. Also random fragment 2 and LCR fragment 6 show a weak gel shift.

	Fragment from	Fragment to	Predicted motifs on fragment	Tool	Fragment name	CTCF band shift
Predicted binding sites	6519	6693	6589	CTCFBSDB	C+1 6589	none
	3576	3775	3671	CTCFBSDB	C-1 3671	weak
	2917	3096	3025	Essex	Essex+1 3025	strong
	2327	2488	2427, 2397	Essex	Essex+2 2427, Essex-2 2397	weak
	4442	4611	4534	Essex	Essex+3 4534	none
	580	745	663	Essex	Essex-1 663	none
	1137	1309	1227	Essex	Essex-3 1227	two bands, weak
	5105	5263	5190	Essex	Essex-4 5190	none
	4689	4863	4775	Essex	Essex-5 4775	strong
	4049	4226	4144	Essex	Essex-6 4144	none
Fragments covering regulatory elements CTCF	6968	7154	none	N/A	LCR P1	none
	7061	7263	none	N/A	LCR P2	none
	7205	7391	none	N/A	LCR P3	none
	7325	7520	none	N/A	LCR P4	none
	7486	7656	none	N/A	LCR P5	none
	7598	7794	none	N/A	LCR P6	weak
	7734	7904	none	N/A	LCR P7	none
	7857	81	none	N/A	LCR P8	none
	55	223	none	N/A	LCR P9	none
	772	950	none	N/A	P890 region	weak
	1031	1201	none	N/A	E2 binding site at 1125	none
Random fragments	7205	7354	none	N/A	Random Frag 1	none
	5475	5639	none	N/A	Random Frag 2	weak
	6638	6790	none	N/A	Random Frag 3	none
	2650	2819	none	N/A	Random Frag 4	none
	3072	3238	none	N/A	Random Frag 5	none
	5856	6010	none	N/A	Random Frag 6	none
	843	983	none	N/A	Random Frag 7	none

Table 23) Results of EMSA tested BPV1 fragments.
Band shifts were classified as strong, medium or weak based on the band shift intensity compared to the positive control on each gel.

3.4.3 Determination of the exact position of the strong CTCF binding site found around nucleotide 3000

With the intention to determine how the position of the CTCF binding site on the fragment influences the gel shift, fragments were designed with the predicted CTCF binding site at either the edge or in the middle of the fragment (Figure 23). EMSA results using these fragments unexpectedly revealed that the predicted binding motif around nucleotide 3000 (E+1) was a false positive. Only the fragment containing E+1 at the 3' edge was able to bind CTCF. Testing additional overlapping fragments revealed that one of the fragments tested resulted in a strong shift even though it did not contain the whole predicted binding motif of E+1. Although the predicted binding site was false, a real binding site must be present in close proximity to the predicted binding site.

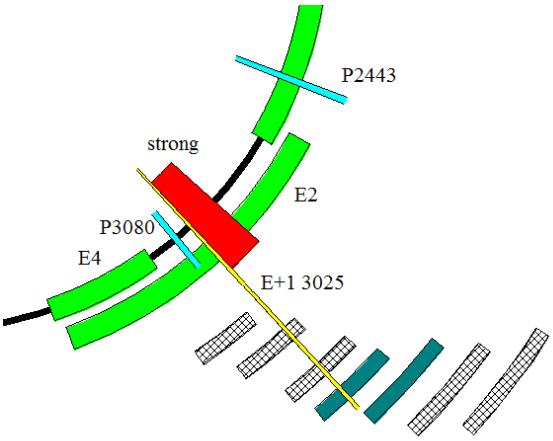


Figure 23) Investigation of the exact position of the CTCF binding site around nucleotide 3000 in BPV1
Only the fragments shown in dark teal bound CTCF in EMSA. Hence the nucleotides those fragments have in common must contain the real CTCF binding site. Note that one of these fragments only contains two out of the 9 nucleotides that make up the predicted binding site E+1 which is shown in yellow. This indicates that motif E+1 is not the correct binding motif of the strong CTCF binding site around nucleotide 3000. Instead the motif must be within the red bar that covers nucleotide 2917 to 3027.

Using several overlapping fragments to determine the exact position of this unpredicted binding site revealed that the size of this site was very large. Despite very long overlaps of the fragment used, only two fragments were able to bind CTCF. These two fragments share 111 base pairs which must contain the real CTCF binding site (nucleotides 2917-3027). There is no match to any of the PWM within this shared sequence. Another study has shown that about 30 % of CTCF binding sites do not fit the consensus 20-mer binding motif of CTCF which might be the case for this particular binding site (Chen *et al.*, 2012a). The shared base pairs of the CTCF binding fragments are indicated by a red bar in Figure 23. The gel shift results of all fragment tested for this experiment are summarised in Table 24 and the primary data can be found in the appendix.

	Fragment		Predicted motifs on fragment	Predicted by	Fragment name	CTCF band shift
	from	to				
Overlapping fragments around nucleotide 3000	2704	2915	none	N/A	No CTCF	none
	2785	2984	none	N/A	Frag CTCF?A	none
	2831	3027	none	N/A	Frag CTCF?B	**strong**
	2917	3096	3025	Essex	E+1 3025 (at 5' edge)	**strong**
	2947	3120	3025	Essex	E+1 3025 (in middle)	none
	2995	3174	3025	Essex	E+1 3025 (at 3' edge)	none

Table 24) EMSA results of fragments tested for CTCF binding around nucleotide 3000 in BPV1

Only two fragments bind CTCF, one of which does not contain the entire predicted binding motif of E+1.

3.4.4 Generation of a CTCF binding map of BPV1

Based on the results from the EMSA it was possible to create a CTCF binding map of the whole BPV1 genome using the programme *DNAman* (Figure 24). The outermost, checked elements of the figure indicate the fragments tested. If a fragment tested positive for CTCF binding it is shown as a blue block on the genome. All predicted binding motifs are shown in yellow and labelled with the tool by which they were predicted and the score of the prediction. If a CTCF binding motif is located within a blue bar it is likely that this is the location where CTCF is binding to the fragment. However since unpredicted binding sites were discovered and many false predictions were revealed it is possible that CTCF binds at a location within the blue bar that is different from the predicted binding site. All promoter sequences are given in teal. The sequence that must contain the CTCF binding site at nucleotide 3000 is shown in red with the yellow prediction of E+1 next to it. The fragments shown in dark teal bind CTCF. The fragment of the LCR that was shown to contain a CTCF binding site also contains a predicted CTCF binding site. However this predicted site is below threshold and therefore coloured purple.

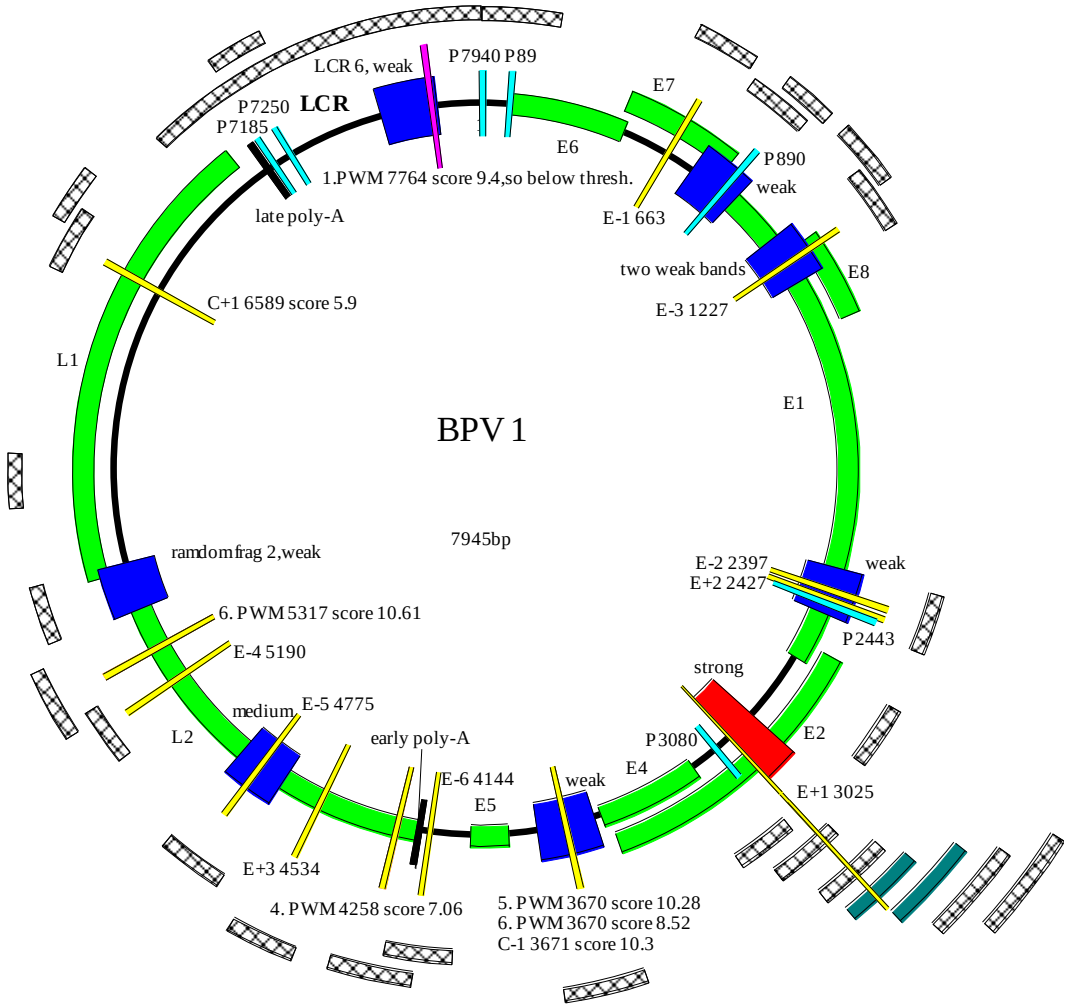


Figure 24) Complete CTCF binding map of the BPV1 genome based on EMSA
Testing coverage is shown as checked fragments in an outer ring around the genome. Fragments tested positive for CTCF binding are shown in blue. Predicted binding sites are indicated with a yellow mark. The unpredicted binding site around nucleotide 3000 is shown in red and the overlapping fragments that bind CTCF are shown in teal.

3.4.5 *In vitro* confirmation of CTCF binding within 7 HPV types based on predictions with CTCFBSDB tool 1.0 and Essex tool

After the confirmation of many CTCF binding sites in the BPV1 genome, it was decided to test the predicted binding sites of the HPV types 18, 16, 16 114/K, 31, 6b, 11, 38 and 8. Many predicted binding sites could be confirmed and some predicted binding sites were tested negative. The high risk types HPV18, 16 and 16 114/K showed a strong shift with the fragment covering the high scoring binding motif around nucleotide 3000. Accordingly, the low risk specific high scoring binding site at nucleotide 5400 tested positive for CTCF binding with a strong shift on the gel. Also every high risk type had a CTCF binding site confirmed in the area around nucleotide 5400; however the shift was weaker and the position of the binding site was not as conserved as in low risk types. In HPV16 and 31 this binding site is located around nucleotide 5200 and in HPV18 it is located at nucleotide 5475.

The beta virus HPV38 did not show a high degree of similarity to high risk or low risk alpha viruses. Only two CTCF binding sites were confirmed on HPV38, one of which is in the LCR. The predicted CTCF binding cluster in beta viruses at nucleotide 3600 did not result in a shift in the EMSA.

Fragments from the HPV8 template DNA did not amplify. The template DNA was digested and the product was run on an agarose gel, resulting in a different banding pattern than expected for the HPV8 construct. Hence HPV8 could not be tested for CTCF binding. All EMSA results are summarised in Table 25. The primary gel shift data used to produce this table can be found in the appendix.

Class	Template	Fragment from	Fragment to	Predicted motifs on fragment	Predicted by	Fragment name	CTCF band shift
High risk	HPV18	2926	3117	2990	CTCFBSDB	HPV18 C+1 2990	strong
		1102	1297	1205	Essex	HPV18 E+1 1205	none
		5381	5577	5475	Essex	HPV18 E-1 5475	medium
		3381	3575	3487	Essex	HPV18 E-2 3487	none
	HPV16	2852	3049	2916	CTCFBSDB	HPV16 C+1 2916	strong
		1212	1405	1283	CTCFBSDB	HPV16 C+2 1283	none
		6426	6600	6512	CTCFBSDB	HPV16 C-1 6512	weak
		6051	6278	6127	Essex	HPV16 E+1 6127	weak
		5000	5207	5119	Essex	cor. HPV16 E-1 5120	weak
	HPV16 114/K	2852	3049	2917	CTCFBSDB	HPV16 C+1 2917	strong
		6051	6278	6128	Essex	HPV16 E+1 6128	weak
		5000	5207	5120	Essex	cor. HPV16 E-1 5120	weak
	HPV31	2357	2531	2413	CTCFBSDB	HPV31 C+1 2413	strong
		1182	1374	1278, 1294	CTCFBSDB, Essex	HPV31 C+2/E+3 1278	none
		2230	2406	2332	CTCFBSDB	HPV31 C+3 2332	none
		5077	5273	5179	Essex	HPV31 E+1 5179	strong
		1029	1200	1093	Essex	HPV31 E+2 1093	none
		804	1008	885	Essex	HPV31 E+4 885	none
Low risk	HPV6b	5317	5515	5425, 5431	CTCFBSDB, Essex	HPV6b C-1 5425 E-1 5431	strong
		4913	5102	5018, 4987	CTCFBSDB, Essex	HPV6b C-2 5018 E+1 4987	none
		1251	1460	1357	Essex	HPV6b E+2 1357	none
		5995	6199	6099	Essex	HPV6b E+3 6099	none
		6179	6382	6264	Essex	HPV6b E-2 6264	medium
		7155	7380	7206, 7257	Essex	HPV6b E-3 7206 E-4 7257	none
		4715	4913	4790	Essex	HPV6b E-5 4790	none
	HPV11	5330	5501	5416, 5422	CTCFBSDB, Essex	HPV11 C-1 5416 E-1 5422	strong
		6243	6428	6311	CTCFBSDB	HPV11 C-2 6311	weak
		1295	1494	1357	Essex	HPV11 E+1 1357	none
		4709	4898	4781	Essex	HPV11 E-2 4781	none
		3930	4153	4058	Essex	HPV11 E-3 4058	medium
Beta PV	HPV38	3436	3666	3571, 3540, 3520, 3487, 3544	CTCFBSDB, Essex	HPV38 C+1/2 and E+2/4/5 3540	none
		1243	1431	1331	CTCFBSDB	HPV38 C+3 1331	medium
		35	219	117	CTCFBSDB	HPV38 C-1 117	medium
		176	363	254	Essex	HPV38 E+1 254	none
		4357	4545	4454	Essex	HPV38 E+3 4454	none
	HPV8	3553	3742	3642, 3650	CTCFBSDB, Essex	HPV8 C-1 /E-2/E-4 3642	N/A
		2262	2481	2383, 2390	CTCFBSDB, Essex	HPV8 C-2 /E-1/E-3 2383	N/A
		3946	4130	4037	Essex	HPV8 C-3 4037	N/A
		7260	7459	7361	Essex	HPV8 E+1 7361	N/A

Table 25) EMSA results of all predictions made with the Essex tool and the CTCFBSDB tool 1.0

This screening revealed many false positives but also showed similarities among many HPV. All high risk types with a predicted, high scoring binding site around nucleotide 3000 showed a strong band shift with the according fragment. HPV31 has no binding site predicted in this region. Additionally fragments around nucleotide 5400 of the low risk viruses tested showed a strong band shift, too.

3.4.6 Testing additional fragments of several HPV types for CTCF binding around nucleotides 3000 and 5200

On the basis of the EMSA data obtained so far it was decided to test for additional fragments on certain viruses. HPV31 was the only high risk virus that did not have a CTCF binding motif predicted around nucleotide 3000. All the other high risk viruses had a CTCF binding site predicted at this position and all fragments containing this binding site resulted in a strong band shift in EMSA. Hence it was decided to design two overlapping fragments for HPV31 that cover the region around nucleotide 3000. Based on the information from overlapping fragments used previously for testing BPV1 binding site E+1, the fragments for HPV31 were designed with a large overlap of 100 nucleotides to make sure that a large binding site would not be disturbed by the possibility of a too short overlap. Also fragments of this region from both low risk viruses were tested for an unpredicted CTCF binding site.

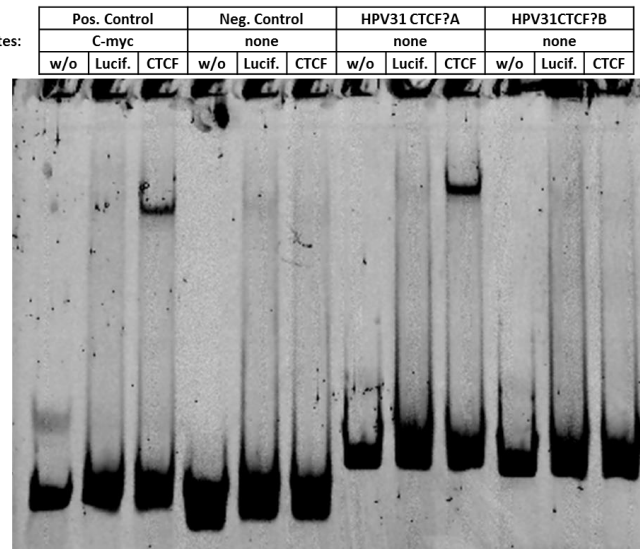


Figure 25) Testing HPV31 fragments for an unpredicted binding site around nucleotide 3000
This image is representative of two independent repeats. One of the fragments shows a strong band shift and confirms the presence of an unpredicted binding site.

Since HPV18 was the only high risk virus that did not have a confirmed CTCF binding site around nucleotide 5200, fragments were designed to test this area for CTCF binding, too. A binding site for nucleotide 5200 at HPV18 could not be confirmed. From all the fragments tested only one of the HPV31 fragments bound CTCF in EMSA as seen in Table 26 and Figure 25. The gel shift observed for this fragment was strong. This shift confirmed the presence of a strong CTCF binding site around nucleotide 3000 in HPV31, revealing that this binding site is present across all high risk viruses tested.

Template	Fragment from	Fragment to	Predicted motifs on fragment	Predicted by	Fragment name	CTCF band shift
HPV18	4947	5155	None	N/A	HPV18 CTCF?A	none
	5045	5253	None	N/A	HPV18 CTCF?B	none
HPV31	2801	3015	None	N/A	HPV31 CTCF?A	**strong**
	2894	3093	None	N/A	HPV31 CTCF?B	none
HPV6b	2801	3007	None	N/A	HPV6b CTCF?A	none
	2887	3101	None	N/A	HPV6b CTCF?B	none
HPV11	2801	3003	None	N/A	HPV11 CTCF?A	none
	2900	3104	None	N/A	HPV11 CTCF?B	none

Table 26) EMSA testing of areas in which many other PV have predicted CTCF binding sites

HPV31 shows a strong shift with one of the fragments tested. Hence all high risk viruses have a confirmed CTCF binding site around nucleotide 3000 that produces a strong shift in EMSA.

3.4.7 *In vitro* confirmation of CTCF binding sites predicted by *Storm*

After extensive testing of binding sites predicted with CTCFBSDB 1.0 and the Essex tool an update of the CTCFBSDB tool to version 2.0 was released. This version of the tool was improved but still flawed as described earlier. Instead of using the new version of the CTCFBSDB tool it was decided to manually predict binding sites using *Storm* and all 6 PWM for CTCF binding that version 2.0 of the CTCFBSDB tool would use. Predictions using *Storm* were carried out by Dr. Jesse Ziebarth (University of Tennessee Health Science Center, USA). Table 27 shows the new fragments used for testing *Storm*-predicted motifs that were not located on fragments tested earlier. Some of the fragments tested showed medium to weak strength shifts. HPV8 was again not tested because no functional template for fragment generation was available. Many newly predicted binding sites were confirmed, most of them showing weak or medium shifts. One fragment covering nucleotides 4440 to 4638 of HPV18 showed two band shifts (Figure 26).

Template	Fragment		Predicted motifs on fragment	Predicted by [score]	Fragment name	CTCF band shift
	from	to				
HPV18	754	943	844	5.PWM [10.0]	HPV18 pred. 844	**none**
	3527	3718	3621	3.PWM [8.2], 5.PWM [14.0], 6.PWM [14.2]	HPV18 pred. 3621	**medium**
	4440	4638	4505, 4538	1.PWM [12.2], 5.PWM [10.5]	HPV18 pred. 4505	**two bands, medium**
	5655	5850	5768	1.PWM [12.9], 3.PWM [9.7], 4.PWM [7.3], 5.PWM [10.1], 6.PWM [12.2]	HPV18 pred. 5768	**medium**
HPV16	6772	6957	6860	1.PWM [12.1], 3.PWM [9.0], 4.PWM [7.7], 6.PWM [9.1]	HPV16 pred. 6860	**weak**
HPV16 114/K	1216	1405	1284	5.PWM [10.2]	114K pred. 1284	none
	6421	6621	6517	1.PWM [13.9], 3.PWM [6.5], 4.PWM [8.6]	114K pred. 6517	**weak**
	6783	6945	6861	1.PWM [12.1], 3.PWM [8.9], 4.PWM [7.7], 6.PWM [9.1]	114K pred. 6861	**weak**
HPV31	534	713	616	4.PWM [8.2]	HPV31 pred. 616	none
	6354	6540	6432	1.PWM [13.5], 3.PWM [7.2]	HPV31 pred. 6432	**medium**
HPV11	4844	5041	4921	4.PWM [7.3]	HPV11 pred. 4921	**weak**
	6544	6738	6636	4.PWM [7.3]	HPV11 pred. 6636	**weak plus smear**
	6872	7074	6980	5.PWM [10.1], 6.PWM [9.3]	HPV11 pred. 6980	none
HPV38	584	791	690	6.PWM [9.2]	HPV38 pred. 690	none
BPV1	4145	4354	4258	4.PWM [7.1]	BPV1 pred. 4258	none
	5232	5422	5317	6.PWM [10.6]	BPV1 pred. 5317	none

Table 27) Gel shift results of the fragments used for testing CTCF binding sites predicted by *Storm*

Only binding sites with a score above threshold are tested. If a *Storm* predicted binding site was located on a fragment tested earlier this fragment is not listed here. A complete summary of all fragments tested and the prediction sites they contain can be found in Table 28 and Table 29.

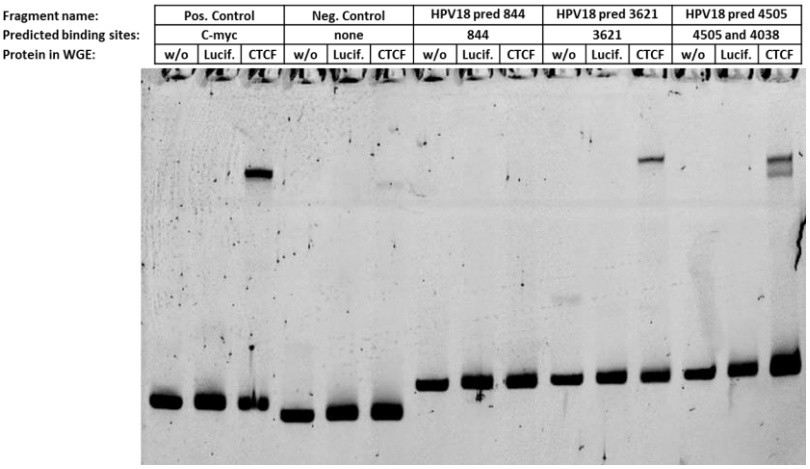


Figure 26) *Storm*-predicted fragments of HPV18 tested for CTCF binding by EMSA

This image is representative of two independent repeats The fragment containing prediction 844 does not bind CTCF. The fragment with prediction 3621 does show a single band of medium strength. Two band where seen with the fragment containing two predicted CTCF binding sites.

On this fragment two independent binding sites were predicted and binding of CTCF to either one of these sites could result in two shifted bands if CTCF binding leads to a distinct migration pattern for each particular site. If this is the case it would imply that the binding of one CTCF protein prevents binding of a second protein on the same fragment because a third band would have been visible on the gel which would be the result of two fragments binding simultaneously to the fragment. However if the binding of CTCF to either one of the predicted binding sites results in an identical migration pattern only one band would be seen for binding to either predicted binding site. In this case the second band would be the result of CTCF proteins binding to each of the two predicted binding sites on the fragment.

3.4.8 Complete summary of all predictions and EMSA results combined

All Fragments tested are summarised in Table 28 and Table 29, it is also indicated whether a fragment contains one or more predicted binding sites and which programme or PWM identified these binding sites. The fragment of HPV31 around nucleotide 3000 that tested positive for CTCF binding now contains a newly predicted binding site. This binding site was found by *Storm* and achieved a high score with multiple PWM. HPV31 is unique in that it contains 3 strong CTCF binding sites. All other HPV tested have only one CTCF binding site that is classified as strong.

The results for the most thoroughly tested papillomavirus, BPV1, are summarised in Table 28. One can see that fragments containing a predicted binding motif were much more often proven to bind CTCF than random fragments or fragments covering important regulatory elements of the PV genome. Hence the prediction tools were helpful in the search for novel CTCF binding sites. However they were not without flaw. Even the approach using *Storm* did not reveal the binding site at nucleotide 3000 of BPV1, despite the fact that the fragment of this region showed a strong shift. Additionally some other fragments that produced medium to weak strength shifts contained no predicted binding motif. However, considering that the EMSA were done *in vitro* in non-physiological binding buffer these results may not fully resemble *in vivo* conditions.

A common CTCF binding pattern among high risk and low risk HPV could be defined. Strong binding sites were only present between nucleotide 2300 and 5500. Binding sites resulting in a weaker shift were more common than binding sites with strong shifts and they accumulated within the ORFs of both late genes. Whereas the biggest difference between high and low risk viruses was the presence of a conserved and strong binding site at nucleotide 3000 or 5400, the binding pattern in the beta virus tested was completely different. Here the only confirmed binding sites were found at nucleotides 116 and 1330, both resulting in a medium band shift. No strong binding sites were found and also accumulation of weaker binding sites in the open reading frames of the late genes was not seen. All primer data and EMSA images can be found in the appendix.

	Fragment		Predicted motifs on fragment	Predicted by [score]	Fragment name	CTCF band shift
	from	to				
Fragments of motifs predicted by *Storm*	4145	4354	4258	4.PWM [7.1]	BPV1 pred. 4258	none
	5232	5422	5317	6.PWM [10.6]	BPV1 pred. 5317	none
Fragments of motifs predicted by by CTCFBSDB 1.0 and Essex tool	580	745	663	Essex	Essex-1 663	none
	1137	1309	1227	Essex	Essex-3 1227	**two bands, weak**
	2327	2488	2427, 2397	Essex	Essex+2 2427, Essex-2 2397	**weak**
	2917	3096	3025	Essex	Essex+1 3025	**strong**
	3576	3775	S 3670, C 3671	5.PWM [10.3], 6.PWM [8.5], CTCFBSDB [10,3]	C-1 3671	**weak**
	4049	4226	4144	Essex	Essex-6 4144	none
	4442	4611	4534	Essex	Essex+3 4534	none
	4689	4863	4775	Essex	Essex-5 4775	**medium**
	5105	5263	5190	Essex	Essex-4 5190	none
	6519	6693	6589	CTCFBSDB [5.9]	C+1 6589	none
Fragments covering regulatory elements	55	223	none	N/A	LCR P9	none
	772	950	none	N/A	P890 region	**medium**
	1031	1201	none	N/A	E2 binding site at 1125	none
	6968	7154	none	N/A	LCR P1	none
	7061	7263	none	N/A	LCR P2	none
	7205	7391	none	N/A	LCR P3	none
	7325	7520	none	N/A	LCR P4	none
	7486	7656	none	N/A	LCR P5	none
	7598	7794	none	N/A	LCR P6	**weak**
	7734	7904	none	N/A	LCR P7	none
	7857	81	none	N/A	LCR P8	none
Random fragments	843	983	none	N/A	Random Frag 7	none
	2650	2819	none	N/A	Random Frag 4	none
	3072	3238	none	N/A	Random Frag 5	none
	5475	5639	none	N/A	Random Frag 2	**weak**
	5856	6010	none	N/A	Random Frag 6	none
	6638	6790	none	N/A	Random Frag 3	none
	7205	7354	none	N/A	Random Frag 1	none
Overlapping fragments at nucleotide 3025	2704	2915	none	N/A	No CTCF	none
	2785	2984	none	N/A	Frag CTCF?A	none
	2831	3027	none	N/A	Frag CTCF?B	**strong**
	2917	3096	3025	Essex	E+1 3025 (at 5' edge)	**strong**
	2947	3120	3025	Essex	E+1 3025 (in middle)	none
	2995	3174	3025	Essex	E+1 3025 (at 3' edge)	none

Table 28) Complete summary of all EMSA-tested fragments of BPV1

In the case of overlapping motifs found by *Storm* only the positions of the motif with the highest score is shown. If motifs of several tools are located on one fragment, the locations are labelled with S for *Storm*, C for CTCFBSDB tool and E for the Essex tool.

Class	Type	Fragment from	Fragment to	Predicted motifs on fragment	Predicted by [score]	Fragment name	CTCF band shift
High risk	HPV18	754	943	844	5.PWM [10.0]	HPV18 pred. 844	none
		1102	1297	1205	Essex	HPV18 E+1 1205	none
		2926	3117	S 2990, C 2990	1.PWM [18.4], 3.PWM [13.2], 4.PWM [13.3], 4.PWM [7.3], 5.PWM [16.6], 6.PWM [23.4], CTCFBSDB	HPV18 C+1 2990	strong
		3381	3575	3487	Essex	HPV18 E-2 3487	none
		3527	3718	3621	3.PWM [8.2], 5.PWM [14.0], 6.PWM [14.2]	HPV18 pred. 3621	medium
		4440	4638	S 4505, S 4538	1.PWM [12.2], 5.PWM [10.5]	HPV18 pred. 4505	two medium bands
		4947	5155	none	N/A	HPV18 CTCF?A	none
		5045	5253	none	N/A	HPV18 CTCF?B	none
		5381	5577	5475	Essex	HPV18 E-1 5475	medium
		5655	5850	5768	1.PWM [12.9], 3.PWM [9.7], 4.PWM [7.3], 5.PWM [10.1], 6.PWM [12.2]	HPV18 pred. 5768	medium
	HPV16	1212	1405	S 1283, C 1283	5.PWM [10.2], CTCFBSDB	HPV16 C+2 1283	none
		2852	3049	S 2916, C 2916	3.PWM [12.2], 5.PWM [9.8], 6.PWM [20.1], CTCFBSDB	HPV16 C+1 2916	strong
		4999	5206	5119	Essex	HPV16 E-1 5119	weak
		6051	6278	6127	Essex	HPV16 E+1 6127	weak
		6426	6600	S 6515, C 6512	1.PWM [13.9], 3.PWM [6.5], 4.PWM [8.6], 6.PWM [9.3], CTCFBSDB	HPV16 C-1 6512	weak
		6772	6957	6860	1.PWM [12.1], 3.PWM [9.0], 4.PWM [7.7], 6.PWM [9.1]	HPV16 pred. 6860	weak
	HPV16 114/K	1216	1405	1284	5.PWM [10.2]	114K pred. 1284	none
		2852	3049	S 2917, C 2917	1.PWM [19.3], 3.PWM [22.4], 4.PWM [16.0], 5.PWM [18.2], 6.PWM [27.6], CTCFBSDB	114K C+1 2917	strong
		5000	5207	5120	Essex	114K E-1 5120	weak
		6052	6279	6128	Essex	114K E+1 6128	weak
		6421	6621	6516	1.PWM [13.9], 3.PWM [6.5], 4.PWM [8.6]	114K pred. 6516	weak
		6783	6945	6861	1.PWM [12.1], 3.PWM [8.9], 4.PWM [7.7], 6.PWM [9.1]	114K pred. 6861	weak
	HPV31	534	713	616	4.PWM [8.2]	HPV31 pred. 616	none
		804	1008	885	Essex	HPV31 E+4 885	none
		1029	1200	1093	Essex	HPV31 E+2 1093	none
		1182	1374	S 1278, C 1278, E 1294	5.PWM [14.0], CTCFBSDB, Essex	HPV31 C+2/E+3 1278	none
		2230	2406	2332	CTCFBSDB	HPV31 C+3 2332	none
		2357	2531	S 2413, C 2413	1.PWM [14.4], 4.PWM [14.7], CTCFBSDB	HPV31 C+1 2413	strong
		2801	3015	2854	1.PWM [19.2], 3.PWM [22.3], 4.PWM [16.0], 4.PWM [9.0], 5.PWM [18.9], 6.PWM [27.5]	HPV31 CTCF?A	strong
		2894	3093	none	N/A	HPV31 CTCF?B	none
		5077	5273	5179	Essex	HPV31 E+1 5179	strong
		6354	6540	6432	1.PWM [13.5], 3.PWM [7.2]	HPV31 pred. 6432	medium

Table 29) Complete summary of all EMSA-tested fragments of high risk papillomaviruses
In the case of overlapping motifs found by *Storm* only the positions of the motif with the highest score is shown. If motifs of several tools are located on one fragment, the locations are labelled with S for *Storm*, C for CTCFBSDB tool and E for the Essex tool.

Class	Type	Fragment from	Fragment to	Predicted motifs on fragment	Predicted by [score]	Fragment name	CTCF band shift
Low risk	HPV6b	1251	1460	1357	Essex	HPV6b E+2 1357	none
		2801	3007	none	N/A	HPV 6b CTCF?A	none
		2887	3101	none	N/A	HPV 6b CTCF?B	none
		4715	4913	4790	Essex	HPV6b E-5 4790	none
		4913	5102	S 5018, C 5018, E 4987	4.PWM [7.5], CTCFBSDB, Essex	HPV6b C-2 5018 E+1 4987	none
		5317	5515	S 5425, C 5425, E 5431	1.PWM [16.9], 3.PWM [17.0], 4.PWM [6.9], 5.PWM [12.9], 6.PWM [12.2], CTCFBSDB, Essex	HPV6b C-1 5425 E-1 5431	strong
		5995	6199	S 6110, E 6099	5.PWM [10.0], Essex	HPV6b E+3 6099	none
		6179	6382	6264	Essex	HPV6b E-2 6264	medium
		7155	7380	E 7206, E 7257	Essex	HPV6b E-3 7206 E-4 7257	none
	HPV11	1295	1494	1357	Essex	HPV11 E+1 1357	none
		2801	3003	none	N/A	HPV 11 CTCF?A	none
		2900	3104	none	N/A	HPV 11 CTCF?B	none
		3930	4153	4058	Essex	HPV11 E-3 4058	medium
		4709	4898	4781	Essex	HPV11 E-2 4781	none
		4844	5041	4921	4.PWM [7.3]	HPV11 pred. 4921	weak
		5330	5501	S 5416, C 5416, E 5422	1.PWM [18.4], 3.PWM [20.5], 4.PWM [11.5], 5.PWM [18.6], 6.PWM [15.8], CTCFBSDB, Essex	HPV11 C-1 5416 E-1 5422	strong
		6243	6428	S 6311, C 6311	4.PWM [13.2], CTCFBSDB	HPV11 C-2 6311	weak
		6544	6738	6636	4.PWM [7.3]	HPV11 pred. 6636	weak, smear
		6872	7074	6980	5.PWM [10.1], 6.PWM [9.3]	HPV11 pred. 6980	none
Beta PV	HPV38	35	219	S 117, C 117	6.PWM [10.9], CTCFBSDB	HPV38 C-1 117	medium
		176	363	254	Essex	HPV38 E+1 254	none
		584	791	690	6.PWM [9.2]	HPV38 pred. 690	none
		1243	1431	S 1331, C 1331	1.PWM [14.9], 5.PWM [15.6], 6.PWM [8.5], CTCFBSDB	HPV38 C+3 1331	medium
		3436	3666	S 3543, S 3573, C 3540, C 3571, E 3520, E 3544 E 3487	1.PWM [13.3], 1.PWM [15.3], 5.PWM [9.2], CTCFBSDB, Essex	HPV38 C+1/2 and E+2/4/5 3540	none
		4357	4545	4454	Essex	HPV38 E+3 4454	none
	HPV8	N/A	N/A	1959	5.PWM [9.1]	N/A	N/A
		2262	2481	S 2387, C 2383, E 2390, E 2346	1.PWM [13.7], 3.PWM [12.1], 4.PWM [11.7], 5.PWM [12.7] CTCFBSDB [12.7], Essex	HPV8 C-2 /E-1/E-3 2380	N/A
		3553	3742	S 3642, C 3642, E 3644 E 3650	3.PWM [12.6]. CTCFBSDB, Essex	HPV8 C-1 /E-2/E-4 3642	N/A
		N/A	N/A	3769	1.PWM [14.3]	N/A	N/A
		3946	4130	4037	CTCFBSDB [6.5]	HPV8 C-3 4037	N/A
		N/A	N/A	5090	1.PWM [11.1], 4.PWM [6.5], 5.PWM [11.3]	N/A	N/A
		N/A	N/A	5280	5.PWM [9.4	N/A	N/A
		N/A	N/A	6172	1.PWM [11.4	N/A	N/A
		7260	7459	7361	Essex	HPV8 E+1 7361	N/A

Table 30) Complete summary of all EMSA-tested fragments of low risk and beta PV

Low risk viruses showed a strong and conserved CTCF binding site around nucleotide 5400. The CTCF binding cluster of beta viruses at nucleotide 3600 was not confirmed. If motifs of several tools are located on one fragment, the locations are labelled with S for *Storm*, C for CTCFBSDB tool and E for the Essex tool.

3.4.9 Creation of EMSA based CTCF binding maps of all HPV tested

All data generated so far have been combined to create CTCF binding maps for all PV types tested using the *DNAman* software. The data shown in the maps on the following pages includes predicted binding sites, testing coverage and EMSA results (Figure 27 to Figure 30). Predicted binding sites are shown as yellow marks on the genome whereas the EMSA testing coverage is shown as checked fragments situated in an outer ring around the genome. If one of those fragments resulted in a shift on the gel this fragment is shown as a blue bar on the genome. Accordingly, if a yellow prediction mark is found on top of a blue bar it is likely that the CTCF protein binds the fragment at the predicted site. However this does not always need to be the case as seen with the overlapping fragments used to test binding site E+1 of BPV1. All prediction marks are labelled with the tool or PWM with which they were predicted. The score of the prediction and its location are also shown. Additionally functional elements like promoters and polyadenylation sites are indicated in the maps if the data was available (source: National Institute of Allergy and Infectious Disease, USA at http://pave.niaid.nih.gov/index.html#transcript) (reviewed by Van Doorslaer et al., 2013). The predicted CTCF binding sites in HPV8 could not be tested by EMSA due to the lack of functional template DNA. Therefore no testing coverage is shown.

Looking at the high risk HPV types only, a common pattern is revealed. In addition to the strong and conserved binding site around nucleotide 3000, every high risk virus contains several more binding sites within the L1 and L2 reading frames. The majority of these binding sites showed a weak or medium strength gel shift. No binding sites were revealed in other regions of high risk papillomaviruses, including the LCR and the first 2000bp of the viral genome. The binding patterns of HPV16 and HPV16 114/K are virtually identical, which is likely to be the result of the high degree of similarity between these viruses. Even most of the predicted binding motifs have an identical sequence as seen in Table 22. Despite the fact that several CTCF binding sites were predicted between nucleotides 500 and 1500 across all high risk and low risk viruses, none of those predicted sites showed a shift in the EMSA experiments, indicating that CTCF does not bind in this region of the HPV genome. On the other hand BPV1 has a confirmed CTCF binding site in this region. In fact, the binding sites across BPV1 are much more distributed across the genome compared HPV types. However BPV1 has a similarity to the high risk HPV types tested which is the strong CTCF binding site around nucleotide 3000.

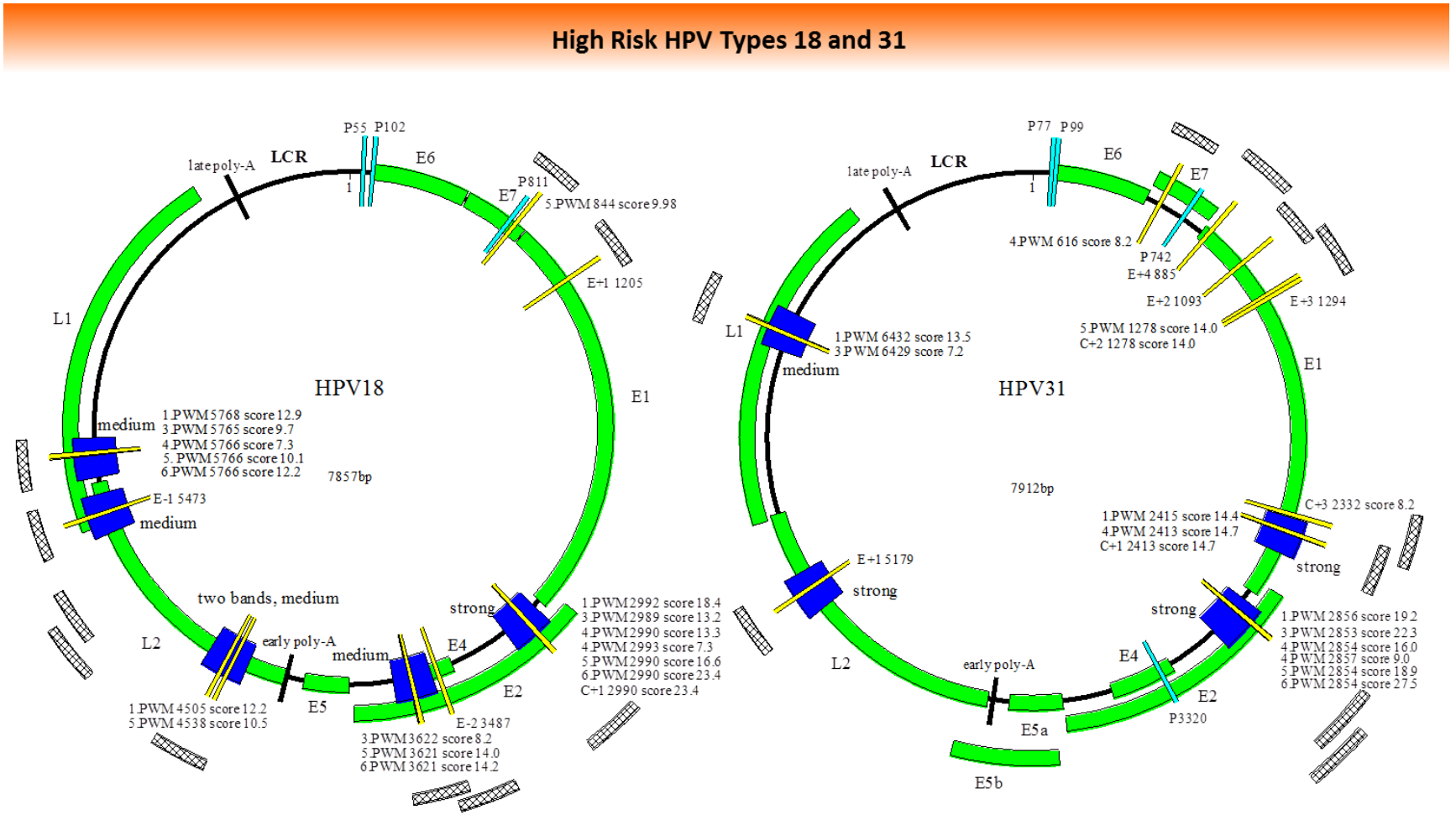


Figure 27) Complete CTCF binding map of the HPV18 and HPV31 genome

Fragments tested are checked and situated in an outer ring around the genome. CTCF binding fragments are shown in blue and predicted binding sites are indicated with a yellow bar. Promoters are shown in teal.

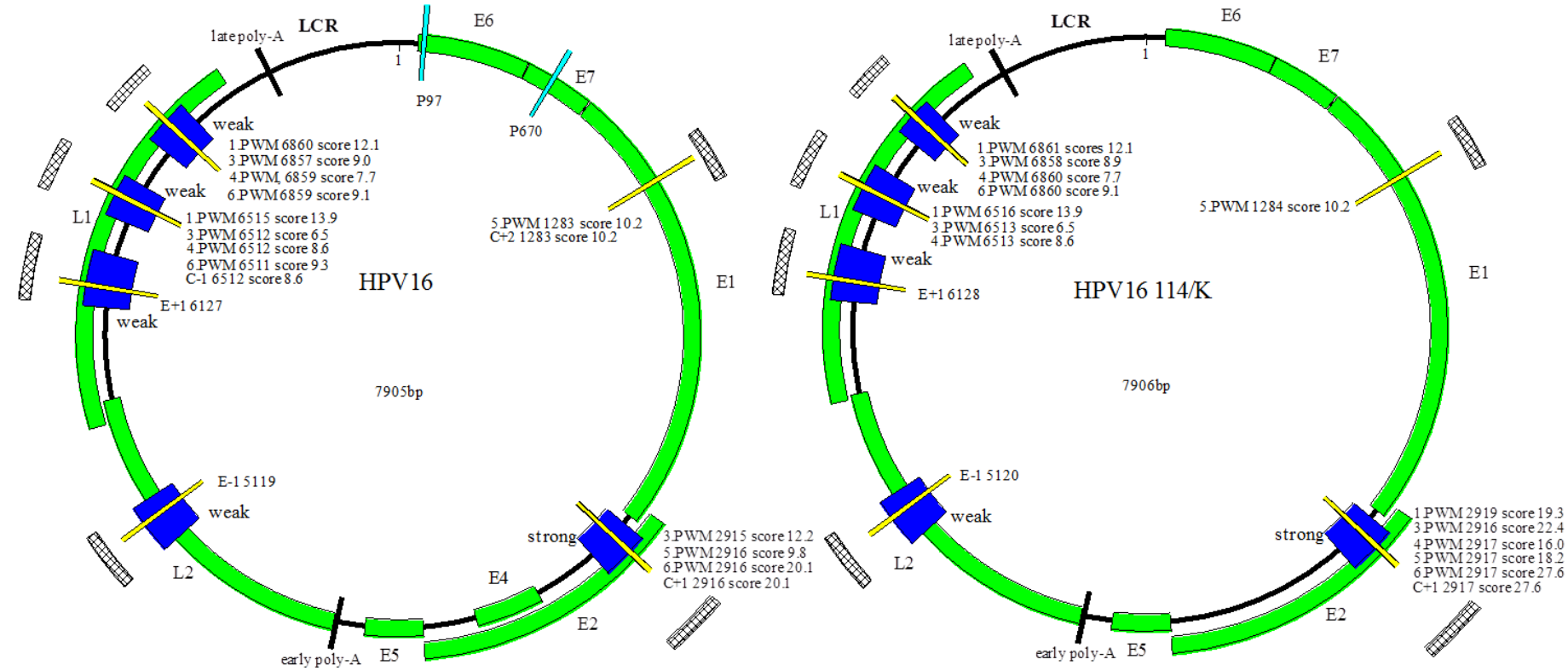


Figure 28) Complete CTCF binding map of the HPV16 and HPV16 114/K genome
Fragments tested are checked and situated in an outer ring around the genome. CTCF binding fragments are shown in blue and predicted binding sites are indicated with a yellow bar. Promoters are shown in teal. No promoter locations for HPV16 114/K were available.

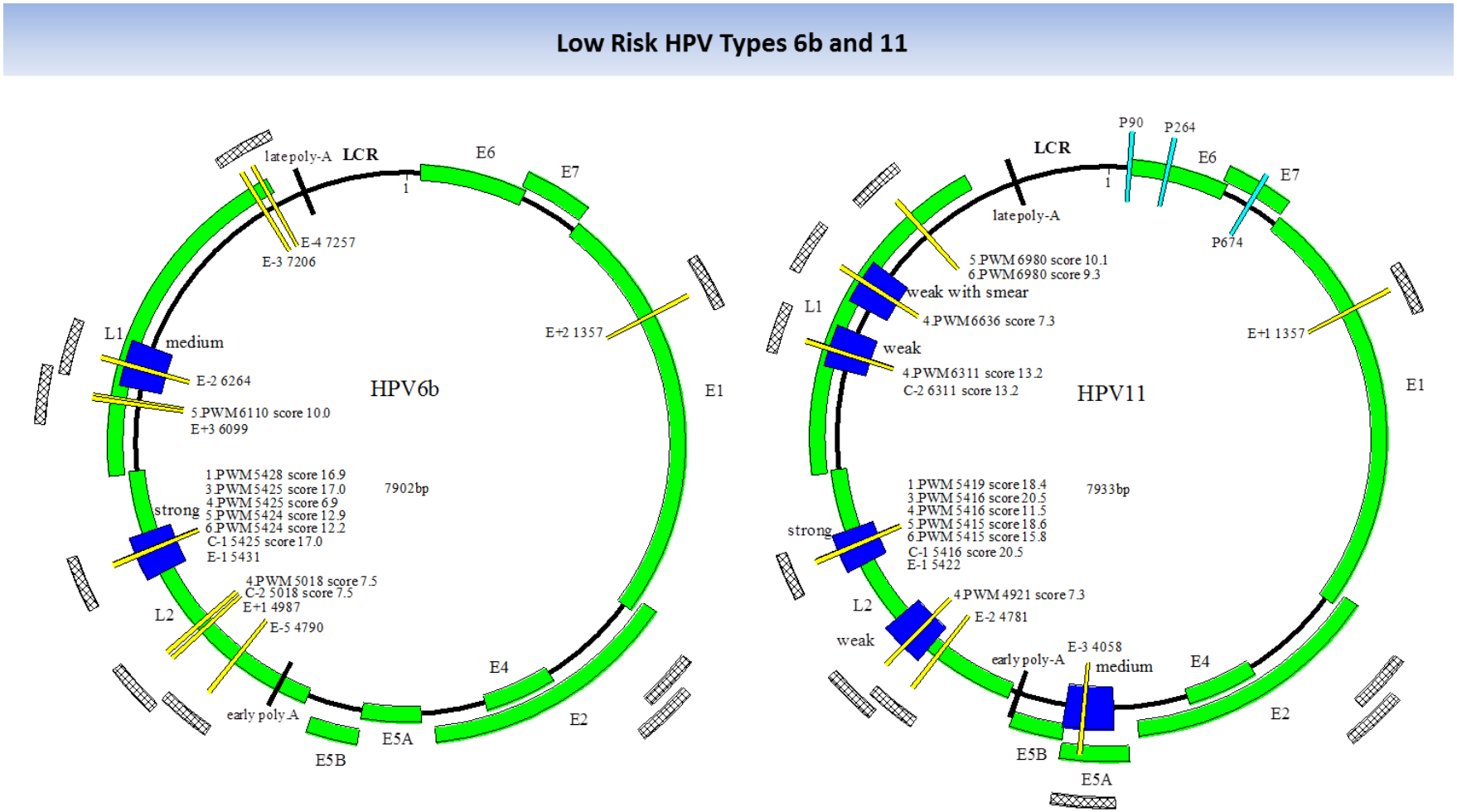


Figure 29) Complete CTCF binding map of the HPV6b and HPV11 genome

Fragments tested are checked and situated in an outer ring around the genome. CTCF binding fragments are shown in blue and predicted binding sites are indicated with a yellow bar. Promoters are shown in teal. No promoter locations for HPV6b were available.

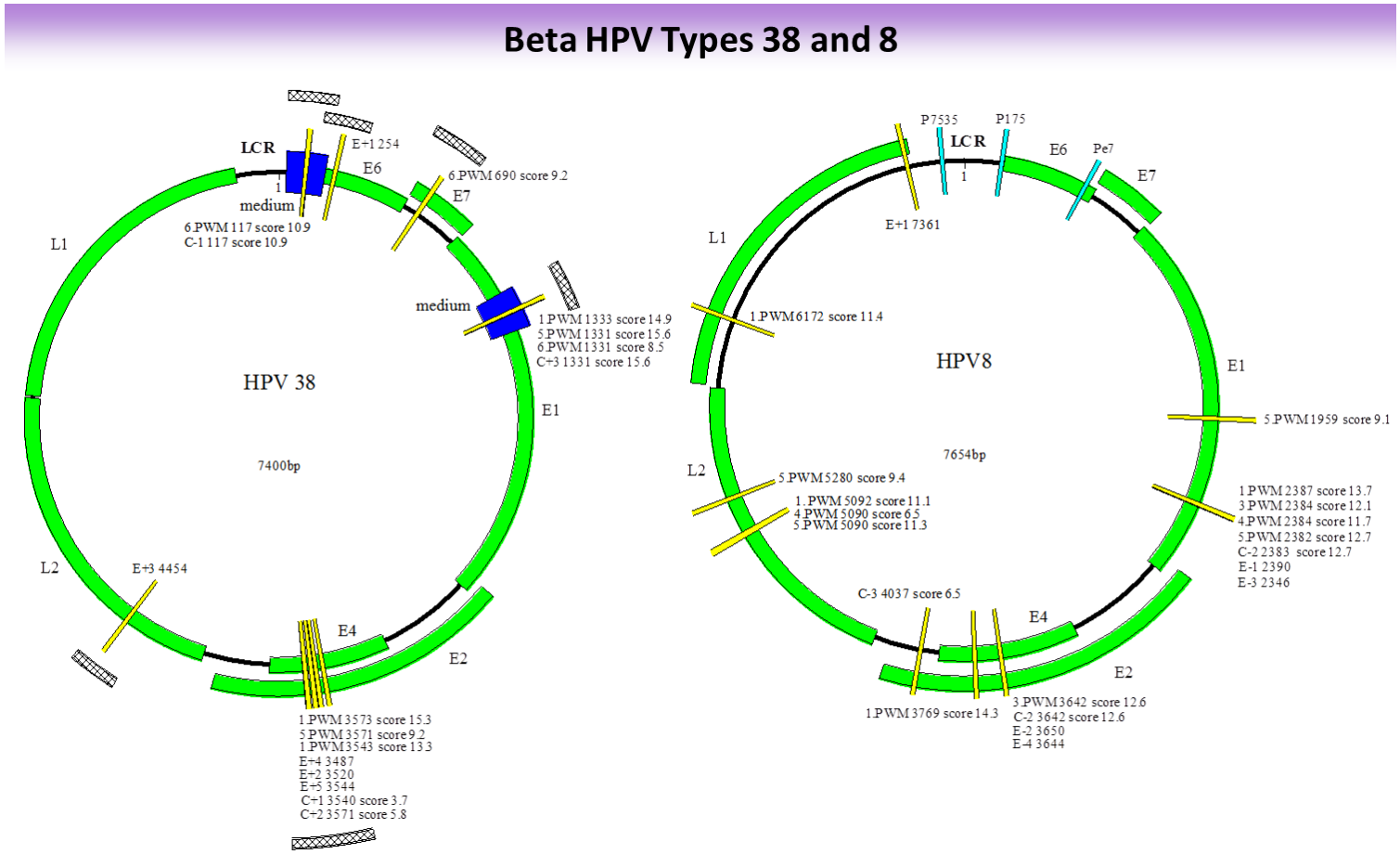


Figure 30) Complete CTCF binding map of the HPV38 genome and predicted CTCF binding sites on HPV8

Fragments tested are checked and situated in an outer ring around the genome. CTCF binding fragments are shown in blue and predicted binding sites are indicated with a yellow bar. Promoters are shown in teal. None of the predicted binding sites of HPV8 have been tested via EMSA since no functional template DNA was available. The promoters of HPV8 were the only functional elements available for these viruses.

3.4.10 Analysis of CTCF binding site prediction tools

With the number of fragments tested it was possible to evaluate the tools used for *in silico* screening of CTCF binding motifs. Table 31 summarises the EMSA results of the fragments designed to test predicted binding site from every tool. Fragments without predicted binding sites are also shown. The table indicates how many of those fragments showed a shift with CTCF protein. Additionally the percentages of fragments showing a gel shift were calculated for each category. The Essex tool predicted 33 binding sites in all the sequences tested, out of which 14 showed a shift on the gel with CTCF protein. This resulted in a ratio of confirmed to predicted CTCF binding sites of 42.4 %. Even though binding site E+1 in BPV1 has been shown not to be the real CTCF binding site on this fragment it was still counted as a confirmed binding site in the table since the other binding sites were not as thoroughly investigated as E+1 to confirm the true location of the binding site. Thus it would have introduced a bias if this fragment was removed from the calculation since the other fragments were not double checked in the same way.

	Number of fragments with predicted binding sites	Fragments binding CTCF *in vitro*	Percentage of confirmed binding sites
Essex tool	33	14	42,4 %
CTCFBSDB tool 1.0	17	11	64,7 %
Storm	33	21	63,6 %
Fragments without a predicted binding site	25	3	12.0 %

Table 31) Comparison of *in silico* prediction tools for CTCF binding motifs
This table does not include the predicted binding sites of HPV8 since this virus was not tested experimentally.

In contrast to the other two tools used, the Essex tool screened for short sequences of 10 bp and since the primary CTCF binding motif is longer than that, the Essex tool was expected to find more false positives than the other tools. The Essex tool has a ratio of confirmed binding sites to predicted binding sites (hit ratio) of 42.4 %. The CTCFBSDB tool 1.0 screened for 4 different PWM and only gave the best match to each PWM while ignoring all lesser matches. This led to a comparably small number of 17 predictions in all sequences screened. However the hit ratio was the highest with 64.7 %. Using *Storm* for screening with 5 different PWM revealed 33 putative binding sites, of which 21 showed a

shift on the gel. This results in a hit ratio of 63.6 % and is therefore only marginally less accurate than the CTCFBSDB tool 1.0 with 64.7 %. The fact that the *Storm* software revealed more total binding sites than the CTCFBSDB tool 1.0, while maintaining a high hit ratio makes this prediction tool superior to the other two prediction tools.

Fragments that were not predicted to bind CTCF with any of the tools used showed gel shifts in 12.0 % of cases. This number is relatively high and may due to the *in vitro* testing of CTCF binding which may be prone to giving false positive results. Alternatively, many binding sites may exist that are not recognized by any of the tools. Since the Essex tool is completely based on *in vitro* EMSA data, it may be possible that the database it uses for screening may also contain many false positive results. If this is the case it would be likely that a lower percentage of EMSA confirmed binding sites of the Essex tool bind CTCF *in vivo* compared to confirmed binding sites predicted with the tools that are based on ChIP data.

Binding sites predicted by *Storm* were split up into predictions for each of the single PWM used by the tool. Table 32 shows how many binding sites above threshold were found with each PWM and how many of the fragments containing these binding sites actually bound CTCF in EMSA. For example, 12 binding sites out of the 33 binding sites found by storm were identified by the 3rd PWM. All 12 fragments bound CTCF, resulting in a hit ratio of 100 %. This is the highest hit ratio of all PWM used even though the third PWM has the lowest score threshold. The 1st PWM has a very similar hit-ratio. Only one out of the 14 fragments containing predicted binding sites from this PWM did not bind CTCF, resulting in a hit ratio of 92.9 %. PWM no. 4 and 6 both predicted 17 binding sites each while also having exactly the same hit ratios at 82.4 %. The fifth PWM exhibits the lowest overall hit ratio with only 64.7 %, even though it has the second highest score threshold.

	Score threshold	Number of fragments	Fragments binding CTCF *in vitro*	Percentage of confirmed binding sites
1.PWM	10,72	14	13	92.9 %
3.PWM	6,45	12	12	100,0 %
4.PWM	6,54	17	14	82,4 %
5.PWM	8,71	17	11	64,7 %
6.PWM	8,51	17	14	82,4 %

Table 32) Comparison of PWM used for *Storm* screening based on EMSA results
PWM 1 and 3 are showing the best hit-ratio at the score threshold used.

Table 33 shows the hit ratio of fragments containing binding motifs predicted by several different PWM. All binding sites recognised by 3 to 5 different PWM showed a shift in EMSA. The ratio of confirmed to predicted binding sites drops down to 66.7 % with motifs predicted by only two PWM and it further drops to 28.6 % for binding sites that were predicted by only one single PWM. Hence the identification of a particular binding motif by several PWM is an important factor that determines the likelihood of this motif for being tested positive for CTCF binding in EMSA.

Number of PWM that identified this site	Number of fragments	Fragments binding CTCF *in vitro*	Percentage of confirmed binding sites
5	6	6	100,0 %
4	3	3	100,0 %
3	4	4	100,0 %
2	6	4	66,7 %
1	14	4	28,6 %

Table 33) Identification of CTCF binding motifs by multiple PWM increases the chance of EMSA confirmation

However the sample size for this analysis was low. This leaves room for criticism. Also a slight bias may have been introduced because fragments from HPV16 and HPV16 114K were very similar and often had identical binding motifs, yet they were still treated as independent fragments in this analysis. Only very few binding sites were predicted by the Essex tool that were also found by either CTCFBSDB 1.0 or *Storm*. Hence binding sites predicted by the Essex tool were not included in Table 33.

3.5 Mutation of a CTCF binding site in the HPV18 genome and *in vivo* confirmation of binding abrogation

3.5.1 Investigating mutations that have the potential to abrogate CTCF binding in EMSA

With the intention to abrogate CTCF binding around nucleotide 3000 and 5400 in HPV18, mutations were investigated that could be introduced into the HPV18 genome. Small FAM-labelled probes were synthesised that contained the CTCF binding site of interest either with or without mutations. It was shown previously that fragments of 60 base pairs are large enough to bind CTCF (Bell *et al.*, 1999). A probe of 60bp containing the HPV18 CTCF binding site at nucleotide 3000 was designed and tested for CTCF binding by EMSA. Another 60bp probe that contained a previously characterised CTCF binding site was used as a positive control for these experiments (Bell *et al.*, 1999). The sequences of these probes can be found in Figure 31 and the gel shifts are shown in Figure 32. In contrast to the positive control probe the HPV18 probe did not bind CTCF in this in vitro reaction. Thus gel shifts with probes containing mutated versions of the binding site were futile.

HPV18 probe:
5'-GGCATACAGACATTAAACCACCAGGTGGTGCCAGCCTATAACATTTCAAAAAGTAAAGCA-3'

Pos. control:
5'-AGGCGCGCCCCCAGGGATGTAATTACGTCCCTCCCCCGCTAGGGGGCAGCGGCGCGCCT-3'

Figure 31) Sequences of pos. control probe and the HPV18 probe containing binding site 3000
In the positive control probe the bases which are crucial for CTCF binding are shown in yellow. The predicted binding site in the HPV18 probe is marked in yellow. This binding site was found on the sense strand by the CTCFBSDB tool 1.0. Considering that it was recently published that a secondary motif downstream of the primary motif can contribute to CTCF binding, the probe was designed to cover 26 base pairs downstream of the predicted binding site to include a potential secondary motif (Schmidt et al., 2012).

A possible reason for not seeing a band shift with the intact CTCF binding site could be that the predicted binding motif was wrong. However the conserved nature of this motif did not support this hypothesis. Furthermore, the fragment screening for the CTCF binding site around nucleotide 3000 in BPV1 revealed that at least some CTCF binding sites need a large stretch of DNA to be able to bind to the fragment. A 60bp probe may not have been enough for CTCF to bind to this particular motif or the CTCF binding site was not positioned optimally within the probe. Based on these possible explanations, it was decided to

continue with site directed mutagenesis of the HPV18 genome and to use fragments amplified from the mutated HPV18 genome to screen for loss of CTCF binding.

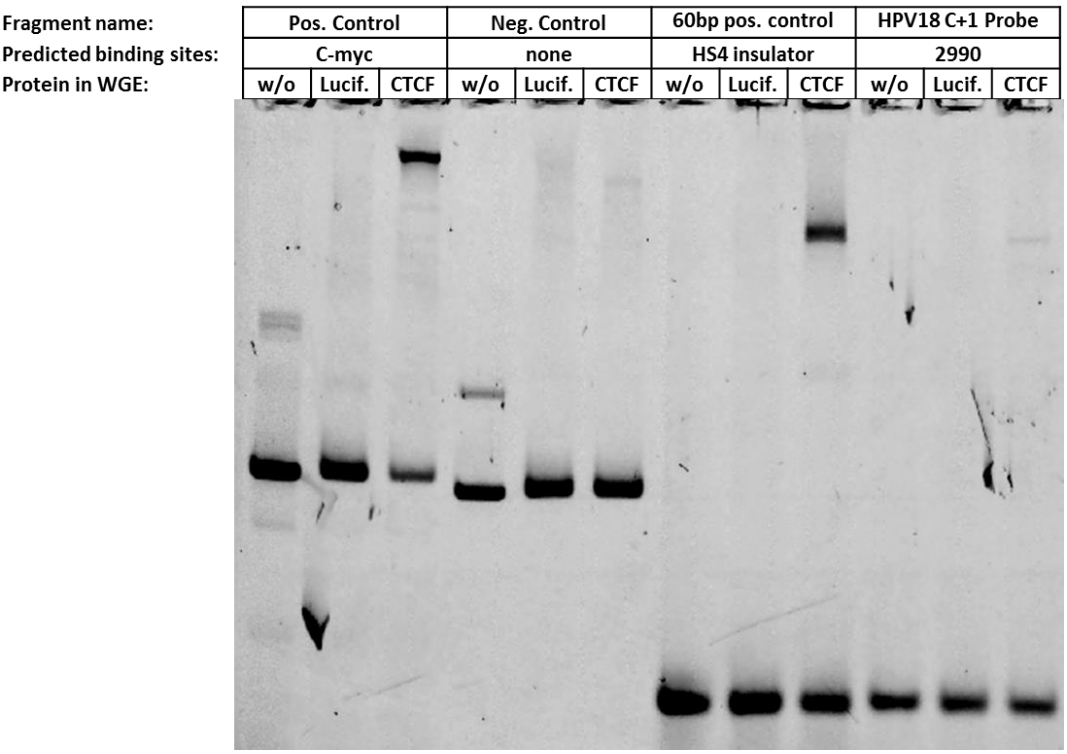
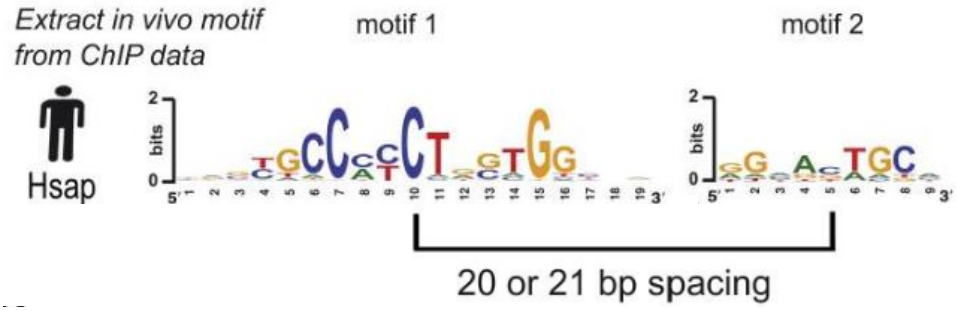


Figure 32) EMSA of 60bp probe for HPV18 binding site 3000
This image is representative of two independent experimental repeats. The control fragments used in previous EMSA are shown in addition to the positive control for the 60bp probe shift. The 60bp positive control probe shows a strong band shift. However the HPV18 probe containing the predicted binding site does not show a shift.

Further analysis of the predicted CTCF binding site gave another indication as to why a shift was not seen with the 60bp probe. The HPV18 probe was designed to include a binding motif found by CTCFBSDB 1.0 with the sequence AAACCACCAGGTGGTGCCAG. This motif was found on the sense strand. Schmidt *et al.* have shown that a secondary motif a few nucleotides downstream of the primary motif can be important for CTCF binding. Therefore, 26 nucleotides downstream of the motif found were included in the probe. At a later date, screening for CTCF binding sites with the PWM defined by Schmidt *et al.* became available using *Storm* or CTCFBSDB 2.0. It was revealed that the primary PWM recognises a CTCF binding site at the same position as CTCFBSDB 1.0 but on the antisense strand instead of the sense strand (motif: CACCACCTGGTGGT). Hence, the secondary motif would be present upstream instead of downstream of this primary motif. In the probe used for EMSA only 10 nucleotides upstream of the primary binding motif were included which means that the secondary binding motif was at the outer edge of the probe what may have resulted in

abrogation of CTCF binding. Figure 33 shows a comparison of the position of primary and secondary binding motifs on the positive control probe and the HPV18 probe.



Positive control sense strand:
5′-AGGCGCGCCCCCAGGGATGTAATTACGTCCCTCCCCCGCTAGGGGGCAGCGGCGCGCCT-3′

Positive control antisense strand:
5′-AGGCGCGCCGCTGCCCCCTAGCGGGGGAGGGACGTAATTACATCCCTGGGGGCGCGCCT-3′

HPV18 probe sense strand:
5′-GGCATACAGACATTAAACCACCAGGTGGTGCCAGCCTATAACATTTCAAAAAGTAAAGCA-3′

HPV18 probe antisense strand:
5′-TGCTTTACTTTTTGAAATGTTATAGGCTGGCACCACCTGGTGGTTTAATGTCTGTATGCC-3′

Figure 33) Probe sequence with binding motifs for the PWM defined by Schmidt *et al*. The primary motif is shown in yellow and the secondary motif is shown in teal. The positive control probe contains both matches to the PWM in the middle of the probe whereas the HPV18 probe has the secondary binding motif on the edge of the sequence. This is likely to be the reason why CTCF did not bind to the probe in EMSA. The PWM was reproduced from Schmidt *et al.* with permission from the copyright holder (Schmidt *et al*., 2012).

3.5.2 Site directed mutagenesis of HPV18 binding sites at nucleotide 3000 and 5400

The gel shift using a 60bp probe containing the HPV18 CTCF binding site at nucleotide 3000 did not show a shift in EMSA experiments. This could either be due to the predicted motif being a false positive, the short length of the probe or the fact that the secondary motif is at the very edge of the probe. The conserved nature of the binding site at nucleotide 3000 among all high risk HPV combined with the EMSA data from larger fragments led to the decision to mutate this binding site in HPV18 regardless the failed gel shift experiments with the 60bp probe. It was decided to look for the binding motifs that fit the PWM defined by Schmidt *et al.* PWM around nucleotide 3000 manually because the expertise for computer based screening for this PWM was not available at the time these experiments were carried out.

The mutations for the abrogation of CTCF binding were designed in a way that the most important bases of the motifs found were altered. In addition to the PWM defined by Schmidt *et al.* the binding site at this position that was identified by CTCFBSDB 1.0 was taken into account.

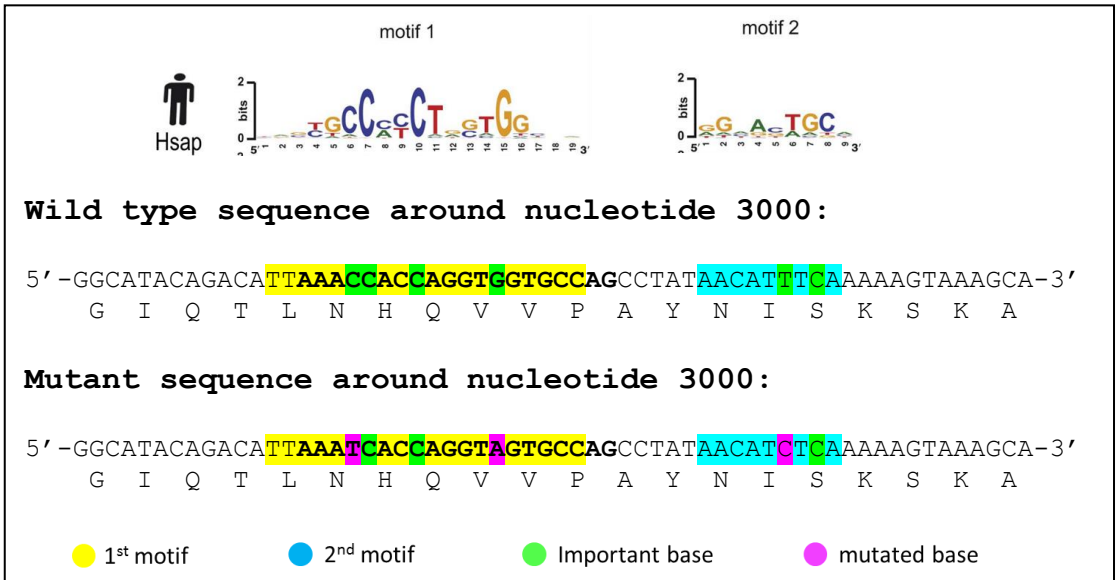


Figure 34) Mutations to abrogate CTCF binding at nucleotide 3000 of HPV18
Bold letters mark the binding motif identified by CCFBSDB 1.0. The motifs of Schmidt *et al.* are shown in yellow and teal. Three important bases for the Schmidt motifs were mutated, two of which also disturbed the binding site identified by CTCFBSDB 1.0. None of the mutations shown changed the amino acid sequence of the E2 protein which is encoded in this region. The PWM was reproduced from Schmidt *et al.* with permission from the copyright holder (Schmidt *et al.*, 2012).

Both motifs were found on the sense strand. Thus it was decided to proceed by mutating the predicted CTCF binding site at nucleotide 3000 on the sense strand of the HPV18 genome by site directed mutagenesis in a way that disturbs both overlaying motifs at once. Figure 34 shows the motif found by CTCFBSDB 1.0 in bold and the motifs matching the PWM defined by Schmidt *et al.* are shown in yellow and teal. The nucleotides to be mutated are marked in purple. All mutations were chosen in a way that did not alter the amino acid sequence of the underlying gene of E2 which is encoded in this region.

The CTCF binding site identified around nucleotide 5400 is highly conserved in low risk viruses and showed a medium shift in HPV18. Therefore this binding site was also mutated. However the binding site around nucleotide 5400 was identified by the Essex tool which does not rate the single bases of the motif by their importance for CTCF binding. Thus the motif defined by Schmidt *et al.* was overlaid to identify important base. Three of these bases were mutated without disturbing the amino acid sequence of the protein encoded in this region. The details about mutating binding site 5400 can be found in Figure 35.

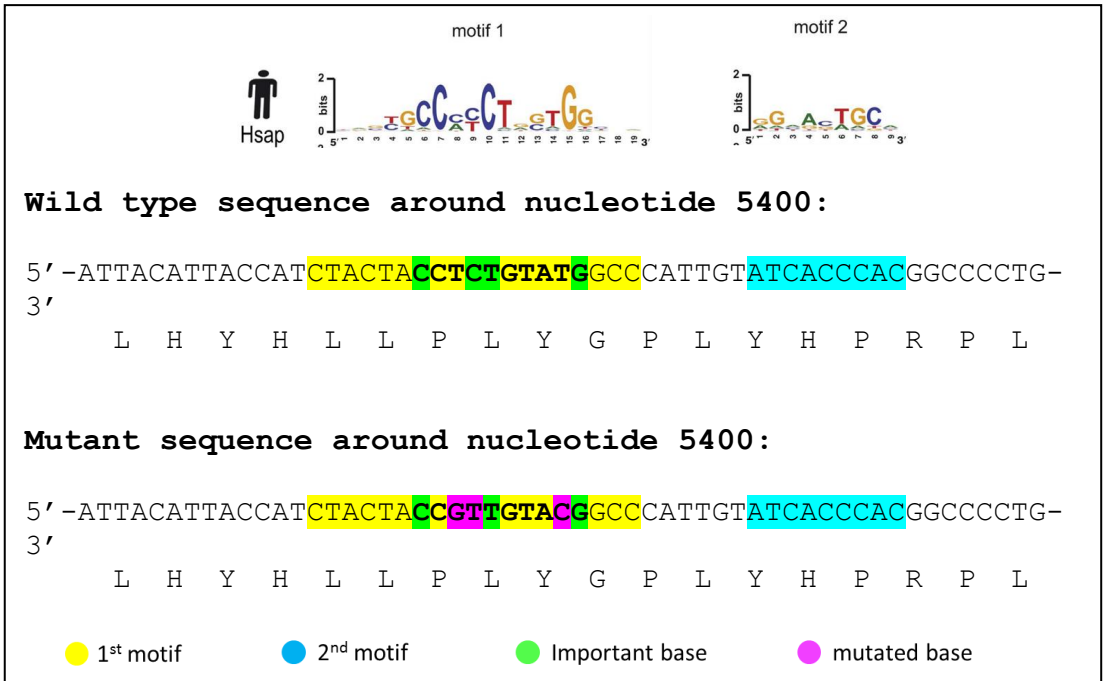


Figure 35) Mutations to abrogate CTCF binding at nucleotide 5400 of HPV18
Three important bases were mutated. None of these mutations changed the amino acid sequence of the underlying protein L2. The PWM was reproduced from Schmidt *et al.* with permission from the copyright holder (Schmidt *et al.*, 2012).

The template used for site directed mutagenesis is the cloning vector pGEM2 containing the complete HPV18 genome (obtained from Frank Stubenrauch, University Hospital Tuebingen, Germany). This construct can be easily amplified in bacteria, which simplifies the procedure of site directed mutagenesis. The HPV18 genome can be excised from the plasmid and re-ligated to re-establish the complete HPV18 genome including the mutations. However pGEM2 combined with the HPV18 genome results in a large plasmid of about 10,700bp which can cause problems with many site directed mutagenesis methods. A specialised kit was used to overcome this issue; the QuikChange II XL Site-Directed Mutagenesis Kit from Agilent Technologies.

The primers used for this technique were carefully designed to fit the requirements of the kit (Figure 36). Mutations had to be located towards the middle of the primer, leaving at least 10 nucleotides at either end that were complementary to the template sequence. Since it was planned to introduce three disparate mutations per CTCF binding site, the mutagenesis primer needed to be long enough to cover all of the mutation sites while still having the 10 complementary nucleotides at either end. The introduction of three mutations at once lowered the melting temperature of the primer but a melting temperature of at least 78°C was suggested by the manufacturers of the site-directed mutagenesis kit. Thus, further elongation of the primer was required to prevent the melting temperature from dropping below 78°C.

Primer for mutating the CTCF binding site at nucleotide 3000

5’-CAGACATTAAATCACCAGGTAGTGCCAGCCTATAACATCTCAAAAAGTAAAG-3’

Primer for mutating the CTCF binding site at nucleotide 5400

5400: 5’-CATTACCATCTACTACCGTTGTACGGCCCATTGTATCACCC-3’

	Binding site 3000	Binding site 5400
Length	52nt	41nt
Tm	78.1°C	78.1°C
Nucleotides before first mutation	11nt	17nt
Nucleotides after first mutation	12nt	17nt
Potential for hairpin formation	none	none
3’ complementarity	none	none

Figure 36) Primers used for mutating the CTCF binding sites around nucleotide 3000 and 5400 in HPV18

The reverse primers used for site directed mutagenesis were the reverse complement of the forward primers shown here.

However using a long primer increased the chance of secondary structure formation and could potentially impair the mutagenesis reaction. Hence each primer was checked for its potential for hairpin formation, 3' complementarity and self-annealing capacity. The forward primers used for the site-directed mutagenesis and their properties are shown in. The reverse primer used in the mutagenesis PCR was the exact reverse complement of the primer containing the mutations. A specialised DNA polymerase suitable for the amplification of large targets was used to synthesise the entire plasmid starting from the mutation primer.

When the expertise for *in silico* screening for the motifs defined by Schmidt *et al.* became available, HPV genome sequences were re-screened and it was revealed that the CTCF binding site at nucleotide 3000 had a second match to the PWM defined by Schmidt *et al.* and this match was located on the antisense strand (Figure 37).

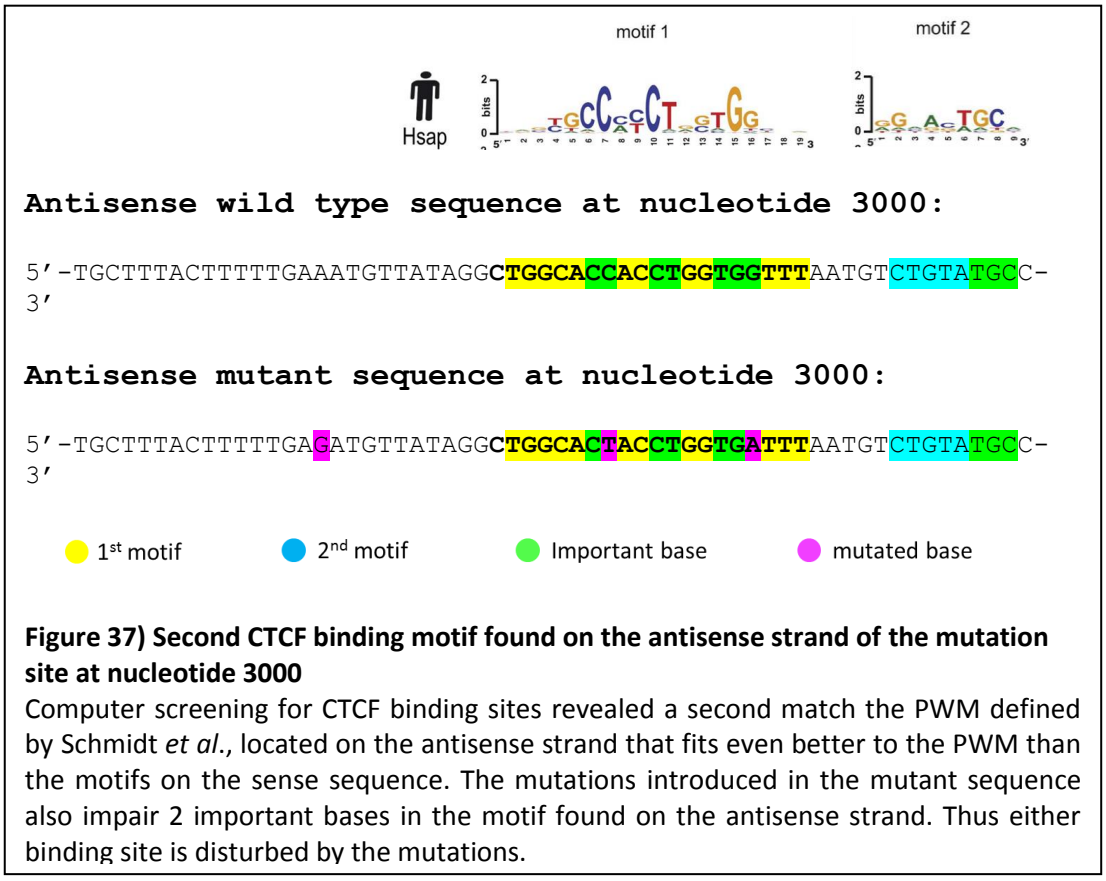


Figure 37) Second CTCF binding motif found on the antisense strand of the mutation site at nucleotide 3000
Computer screening for CTCF binding sites revealed a second match the PWM defined by Schmidt *et al.*, located on the antisense strand that fits even better to the PWM than the motifs on the sense sequence. The mutations introduced in the mutant sequence also impair 2 important bases in the motif found on the antisense strand. Thus either binding site is disturbed by the mutations.

Since the match on the antisense strand was an even better fit to the PWM than the one on the sense strand, it was likely that the motif on the antisense strand was the real binding motif for CTCF. Figure 37 shows the antisense sequence of the mutation site and the location of the newly discovered match to the PWM defined by Schmidt *et al.* It is also shown how the mutations introduced affect this second match to the PWM. The mutations

disturbed two important bases of the motif on the antisense strand. Therefore it was likely that CTCF binding to the motif on the antisense strand was also abrogated. The mutation site at nucleotide 5400 was also screened for the PWM defined by Schmidt *et al.* and no additional binding site was found here.

In silico screening for potential transcription factor binding sites covering both mutation sites was done to account for any possible loss or gain of transcription factor binding other than CTCF. This screening was performed using MATCH™ from Biobase (available at http://www.gene-regulation.com/cgi-bin/pub/programs/match/bin/match.cgi) (Kel et al., 2003). A balanced setting was used for screening that reduced false positive results and false negative results at once. The sequences screened are the ones shown in the figures for mutating both CTCF binding sites (Figure 34 and Figure 35). Every potential transcription factor binding site is given a score that shows how well the potential binding site matches to the core motif of the PWM used for a particular transcription factor. These scores are calculated with a different method than the one used for determining the scores of CTCF binding sites. Thus, these scores are not comparable with one another. The maximum score that can be given my MATCH™ is 1.000. All potential transcription factor binding sites across the mutation sites that were found by MATCH™ are summarised in Table 34.

Transcription factor	Wild type 3000 [core score]	Mutated 3000 [core score]	Wild type 5400 [core score]	Mutated 5400 [core score]
MyoD	**1.000**	-	-	-
Oct1	**0.964**	**0.982**	-	-
GATA3	-	**0.955**	-	-
SOX9	-	-	**1.000**	**1.000**
Myb	-	-	-	**1.000**

Table 34) Screening for transcription factor binding sites across the mutation sites
The mutation of the CTCF binding site around nucleotide 3000 disrupts the potential binding site for MyoD, increases the affinity for Oct1 and creates a binding site for GATA3. The mutation of the CTCF binding site around nucleotide 5400 does not change the affinity for SOX1, but a new binding site for Myb is created.

The wild type CTCF binding site around nucleotide 3000 binds Myogenic Differentiation 1 (MyoD) with a score of 1.000 and Octamer Transcription Factor 1 (Oct1) with a score of 0.964. In the mutant of CTCF binding site 3000, the binding site for MyoD is disturbed and the score for Oct1 is slightly increased to 0.982. MyoD is not expressed in the HPV model cell line Human Foreskin Keratinocytes (HFK), so the loss of the binding site for MyoD may

have no effect on HPV gene expression (Woodman, University of Birmingham, personal communication). Also a putative new binding site for GATA binding protein 3 (GATA3) with a score of 0.955 is created in the mutant. This transcription factor is expressed in HFK, so potential GATA3 binding and the slightly increased affinity of the Oct1 binding site could account for functional differences in the mutant in addition to the potential loss of CTCF.

The mutations of CTCF binding site 5400 do not change the affinity for SRY-box 9 (SOX9) which still scores 1.000 in the mutant. However a new potential binding site for myeloblastosis viral oncogene homolog (Myb) with a score of 1.000 is formed. Myb can activate or repress gene transcription and may have an influence on viral gene expression (Vargova et al., 2011).

Having all the mutagenesis primers available for mutating both CTCF binding sites, three different mutants were planned; first, a mutant of binding site 3000; second, a mutant of binding site 5400 and; third, a double mutant of both binding sites. From here on these mutants are referred to as C3 for mutant 3000, C5 for mutant 5400 and C3C5 for the double mutant. Following the PCR-based site directed mutagenesis reaction, the remaining template DNA was degraded using the methylation specific nuclease *Dpn*I. The PCR product was then transformed into *E.coli* XL10-Gold ultracompetent cells, spread on an agar plate and single colonies were picked to grow overnight cultures. Plasmids from these overnight cultures were isolated using a Qiagen mini prep kit. The regions around the mutation sites of the isolated plasmids were sequenced and many clones were shown to contain the desired mutations. From each of the three mutants a particular clone was chosen to be sequenced in full. Only the clone for C5 showed a single point mutation at nucleotide 4360 (G→A). However this mutation is silent and was shown to occur naturally in other HPV18 isolates like CU11 (accession number GenBank: GQ180787.1). Therefore the plasmid with this mutation was still used for future experiments.

3.5.3 Confirmation of CTCF binding abrogation at mutated sites within the HPV18 genome and re-construction of the viral episome

Before transfecting the mutated genomes into cells, the CTCF binding abrogation in the mutant genomes needed to be confirmed *in vitro* using EMSA. The same primer pair was used to amplify the binding site-containing fragment from wild type (WT), C3 and C3C5 genomes. Those fragments were tested by EMSA as described previously and the results are shown in Figure 38 and Figure 39. The C3 mutations resulted in a near complete abrogation of CTCF binding. Quantification of the band shifts revealed that the C3 mutant only showed 2.6 % of the CTCF binding of the wild type. This mutant was therefore used for further experiments. However the fragment containing the mutated C5 binding site still bound CTCF as efficiently as the WT fragment (Figure 38 and Figure 39).

Either the mutations introduced did not affect CTCF binding or the actual CTCF binding site on this fragment was not situated at the predicted binding motif and was therefore not altered by the mutations. Hence the C5 and the C3C5 mutants were not used for transfection into primary human foreskin keratinocytes.

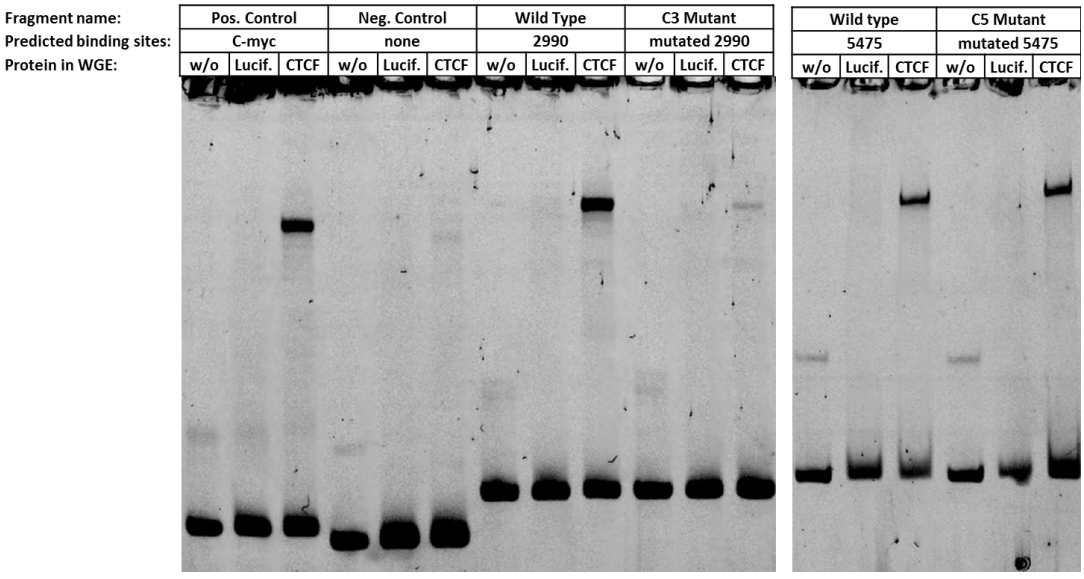


Figure 38) Testing mutations at nucleotide 3000 and 5400 for CTCF binding abrogation
The EMSA image shown here is representative of three independent experimental repeats. The wild type shows a strong band with fragment 2990 whereas the shift in the C3 mutant is almost completely gone. Binding site 5475 does not show a change in band shift with the mutant fragment. The mutations introduced here did not affect CTCF binding.

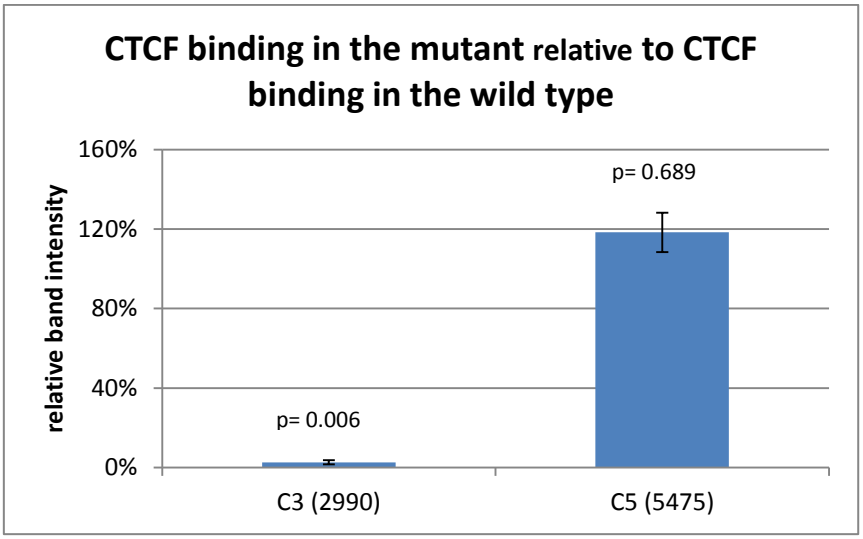


Figure 39) Quantification of CTCF band shifts for wild type and mutants
The bands of all three independent repeat experiments were quantified and normalised against the positive control. The p values were calculated using the double sided t-test against the null hypothesis that the CTCF binding in the mutant was the same as in the wild type. The difference in CTCF binding at 2990 is significant. The C3 mutant only shows 2.6 % of the CTCF binding seen in the wild type. No significant difference in CTCF binding was found when comparing wild type and C5 mutant.

Before the C3 mutant and wild type HPV18 genomes could be transfected into primary cells, the viral genome needed to be excised from the plasmid backbone using *Eco*RI followed by re-ligation to re-establish an uninterrupted HPV18 episome. The ligation was performed using T4 ligase overnight at 16 °C. The DNA from each of the single steps for reconstruction was run on an agarose gel to check the banding pattern of digested and ligated DNA as seen in Figure 40. The sample digested with *Eco*RI showed two bands as expected. The vector pGEM2 is 2869 bp and the HPV18 genome is 7857 bp. Re-ligation resulted in random ligation products of these two fragments, including the ligation product that was a re-constructed ring of the HPV18 DNA only. In this form the viral episome could be transfected into primary human foreskin keratinocytes (HFK). No purification of the ligation products was necessary since only re-ligated, complete HPV18 episomes can be stably maintained by the host cell.

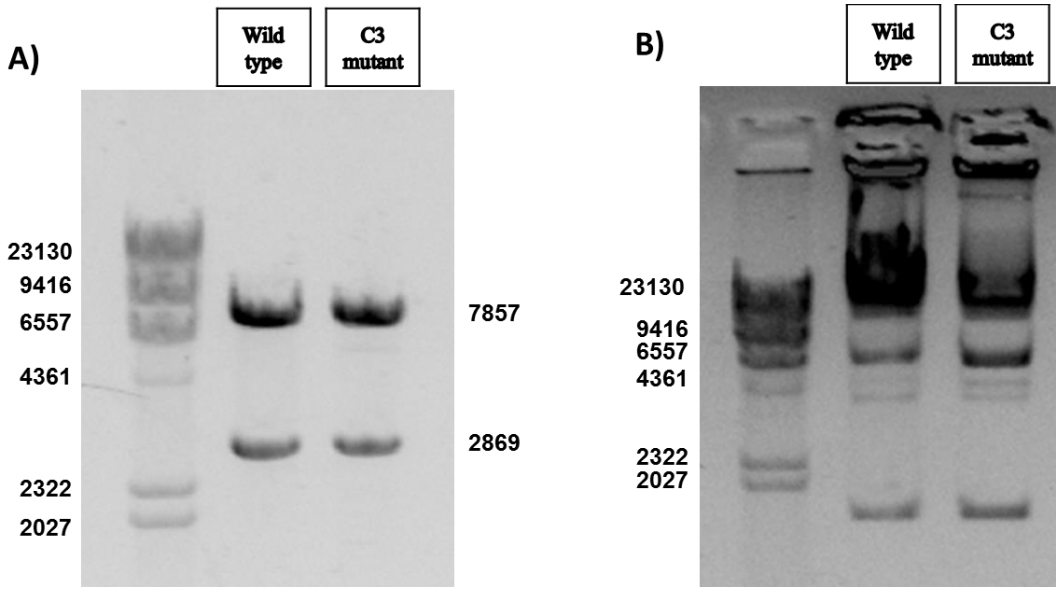


Figure 40) Digestion and re-ligation of plasmids containing WT and C3 mutant HPV18 DNA

Both images are representative of two independent technical repeats.

A) Wild type and mutant plasmid were digested with *Eco*RI overnight. The linearised viral episome is found at 7857bp and the linearised pGEM2 is localised at 2869bp. **B)** Random ligation of all fragments present resulted in multiple bands. Also the viral genome was re-ligated to from the circularised episome.

3.5.4 Creation of cell lines of human foreskin keratinocytes that stably express wild type or mutant HPV18

In collaboration with Dr. Sally Roberts at the University of Birmingham, UK, re-ligated viral episomes were transfected by Dr. Jo Parish into primary human foreskin keratinocytes from two different donors. Cells from two donors were chosen to reduce donor-specific bias. From the samples of each donor two cell lines were created out of which one contained the wild type virus (WT) and the other one the C3 mutant, resulting in overall 4 cells lines. Viral episomes were co-transfected together with a neomycin marker plasmid into sub-confluent primary HFK with subsequent selection using G418. Viral oncogenes transcribed from these HPV18 episomes immortalised the host cells. Continuous culturing led to senescence in cells without viral episome, leaving a continuous cell line of viral genome-containing cells (Hubert and Laimins, 2002, Wilson and Laimins, 2005).

A fragment across the mutation site was amplified by PCR from whole DNA extracted from each cell line and sent for sequencing. It was shown that all cell lines contained the virus

intended, i.e. there was no cross contamination found in any of the cell lines (the alignment is shown in the appendix, page 209). The cell lines from the first donor were named Georgie WT (GWT) and Georgie C3 (GC3). Accordingly the cell lines of the second donor were Clonetics WT (CWT) and Clonetics C3 (CC3).

Since PV episomes can, by chance, integrate into host DNA it was vital to check that the episomes introduced were maintained episomally instead of being integrated. To do so a southern blot was prepared by Dr. Jo Parish. Whole cell DNA from each cell line was sheared and then digested with two different restriction enzymes. *Eco*RI cuts the viral episome once, resulting in linearised DNA of 8kb as seen in Figure 41 and Figure 42. The enzyme *Bgl*II does not cut the viral episome but it frequently cuts host DNA. If the viral episome is maintained episomally two bands are seen on the southern blot, one for the open circle episome and one for the supercoiled episome.

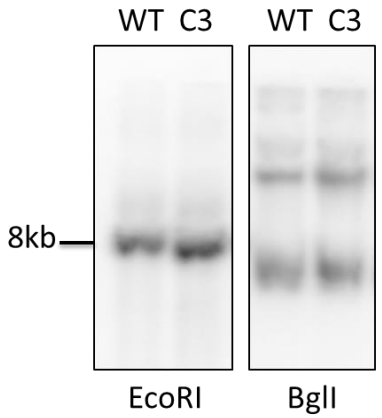


Figure 41) Southern blot of WT and C3 genomes isolated from Georgie HFK cells

This image is representative of two independent experimental repeats (cropped section of Figure 42). *Eco*RI cuts the HPV18 genome once, resulting in a single band of approximately 8kb. *Bgl*II does not cut the HPV18 genome but cuts the host genome. The bands seen following *Bgl*II digestion resemble HPV18 episomes in supercoiled and open circle form. This confirms that the HPV genomes are episomally maintained. This data has been kindly provided by Dr. Jo Parish, University of Birmingham.

Digestion of episomal DNA with *Eco*RI results in linearised DNA shown as a single band of 8 kb on the southern blot. Also digestion with *Bgl*II resulted in the expected banding pattern of supercoiled episomal DNA. Thus the presence of episomal HPV DNA was confirmed. If the viral episome is integrated, digestion with either *Eco*RI or *Bgl*II cuts the viral genomes out of the host DNA, resulting in linearised viral episomes with host DNA attached to either end of it, theoretically resulting in a different banding pattern (Dall *et al.*, 2008). However single integrants would be below the detection limit of this experiment and could possibly co-exist with episomal HPV DNA.

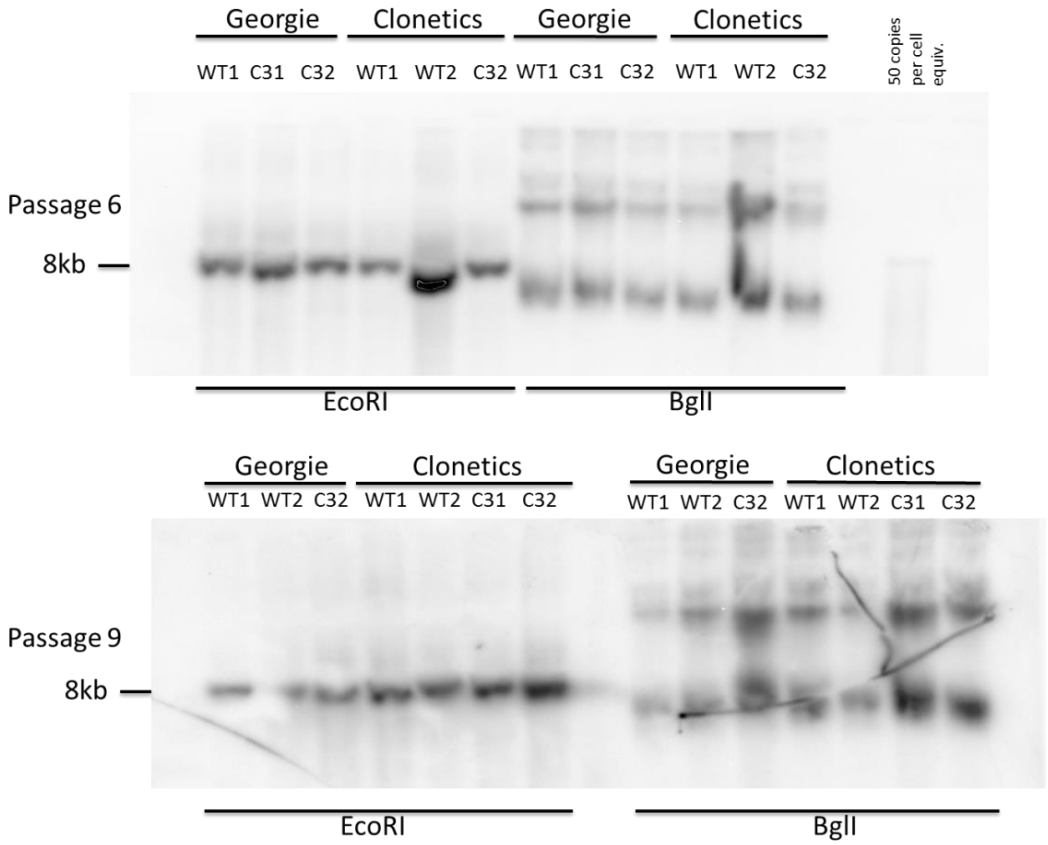


Figure 42) Southern blot for HPV18 DNA from each of the cell lines created
The images are representative of two independent experimental repeats. Some cell lines have been transfected in duplicate and also these duplicates are tested. The single cutter *Eco*RI linearises the viral episome so that it migrates at 8kb. All cell lines show episomal HPV18 DNA. Digestion with *Bgl*II only cuts host DNA and leaves the viral episome intact. Here open circle and supercoiled episome DNA can be seen. This data has been kindly provided by Dr. Jo Parish, University of Birmingham.

3.6 Analysis of cell lines maintaining wild type or mutant HPV18

3.6.1 Collaboration project

The data presented in this chapter was generated in collaboration with the University of Birmingham and the University of Cambridge. The exchange of expertise was essential for the success of this project and collaboration with the University of Birmingham made it possible to work on HPV infected human cells despite the lack of ethical approval for such work at St. Andrews University.

3.6.2 ChIP testing of the viral CTCF binding pattern in cell lines maintaining wild type and C3 mutant HPV18

All experiments presented in the previous sections confirmed CTCF binding sites *in vitro*. However the CTCF binding pattern based on *in vitro* results may not fully resemble *in vivo* conditions, especially considering the unexpected abundance of CTCF binding sites found using EMSA. Thus it was decided to test all predicted CTCF binding sites again using chromatin immunoprecipitation (ChIP) on the cell lines maintaining wild type and C3 mutant HPV18. The ChIP experiments were carried out using the ChIP-IT® Express Enzymatic Kit from Active Motif. Since viral gene expression changes upon differentiation of the host cell the CTCF binding pattern in the viral genome may change as well. However it was decided to exclusively use undifferentiated cells for ChIP since differentiated and therefore keratinous cells were expected to cause major problems with the sample preparation. Undifferentiated cells were formaldehyde fixed, harvested by scraping and dounced with a dounce homogenizer to generate a nuclear preparation. Figure 43 shows HPV18 maintaining HFK cells before and after douncing. These nuclear preparations were treated with a cocktail of restriction enzymes that enter the nuclei through nuclear pores and cut the DNA into fragments. Since the DNA is maintained in nucleosomes, enzymatic digestion with this cocktail of randomly cutting enzymes was only possible in-between nucleosomes. This resulted in shearing of the DNA into fragments that are multiples of 150bp. The enzymatic digestion was timed and stopped after the desired degree of fragmentation was reached. A fragmentation pattern was chosen of which the majority of fragments were between 150 and 450 base pairs in size as seen in Figure 44.

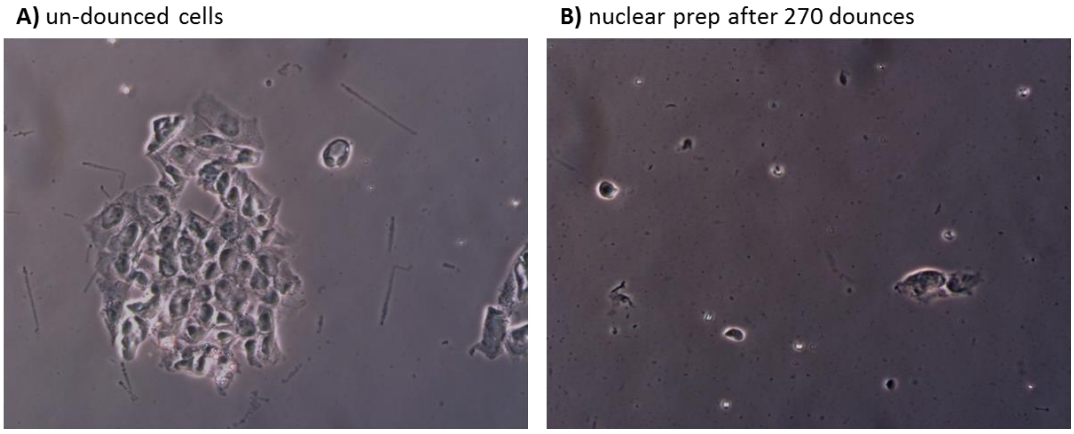


Figure 43) Sample preparation of crosslinked HFK cells maintaining HPV18
Both images are representative of three independent experimental repeats**.**
A) Crosslinked cells harvested by scraping. **B)** Same cells after 270 dounces. The cell membrane is disrupted and the nuclei are released.

Immunoprecipitation was performed by incubating these fragments with antibodies targeting either CTCF or FLAG (negative control) followed by precipitation of protein-DNA complexes bound to these antibodies using protein A coated magnetic beads. Precipitated fragments were then eluted, crosslinks were reversed and the sample was Proteinase K digested before purifying it using a PCR clean-up kit. Subsequently, the samples were analysed by qPCR. Primers used to check predicted CTCF binding sites amplified a region that either contained the predicted binding site or had a predicted binding site in its close vicinity. Also fragments that did not have a predicted CTCF binding site in their vicinity were tested to act as negative control and to give a more complete coverage of the HPV18 genome. All fragments used to test various predicted binding sites are shown in Figure 45 and Table 35.

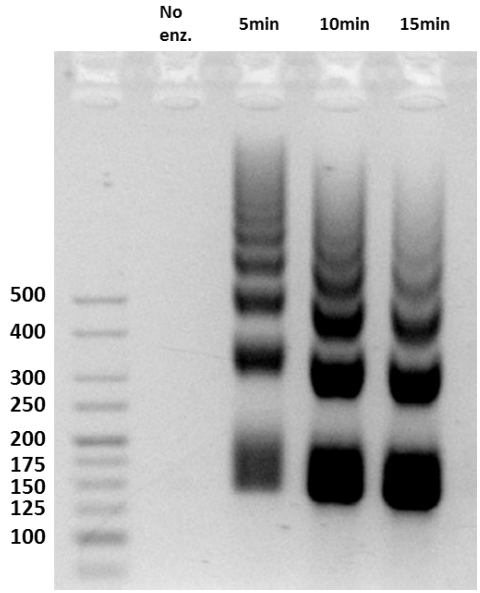


Figure 44) Enzymatic fragmentation of crosslinked DNA-protein complexes
This image is representative of three independent experimental repeats. The fragmentation reaction was stopped at different time points. Digesting the DNA for 10 minutes resulted in a fragmentation pattern appropriate for ChIP.

Fragment name	Fragment start	Fragment end	Predicted binding sites tested
844	789	890	5.PWM 844 [9.98]
1205	1165	1256	E+1 1205
2200	2167	2251	None
2990	2968	3069	6.PWM 2990 [23.4]
3550	3519	3620	E-2 3487 and 6.PWM 3621 [14.2]
4505	4442	4536	1.PWM 4505 [12.2] and 5.PWM 4538 [10.5]
5475	5405	5506	E-1 5475
5768	5712	5811	1.PWM 5768 [12.9]
6400	6371	6462	none
7800	7746	7845	none

Table 35) Fragments used for qPCR analysis of ChIP samples
Fragments across the whole genome are tested.

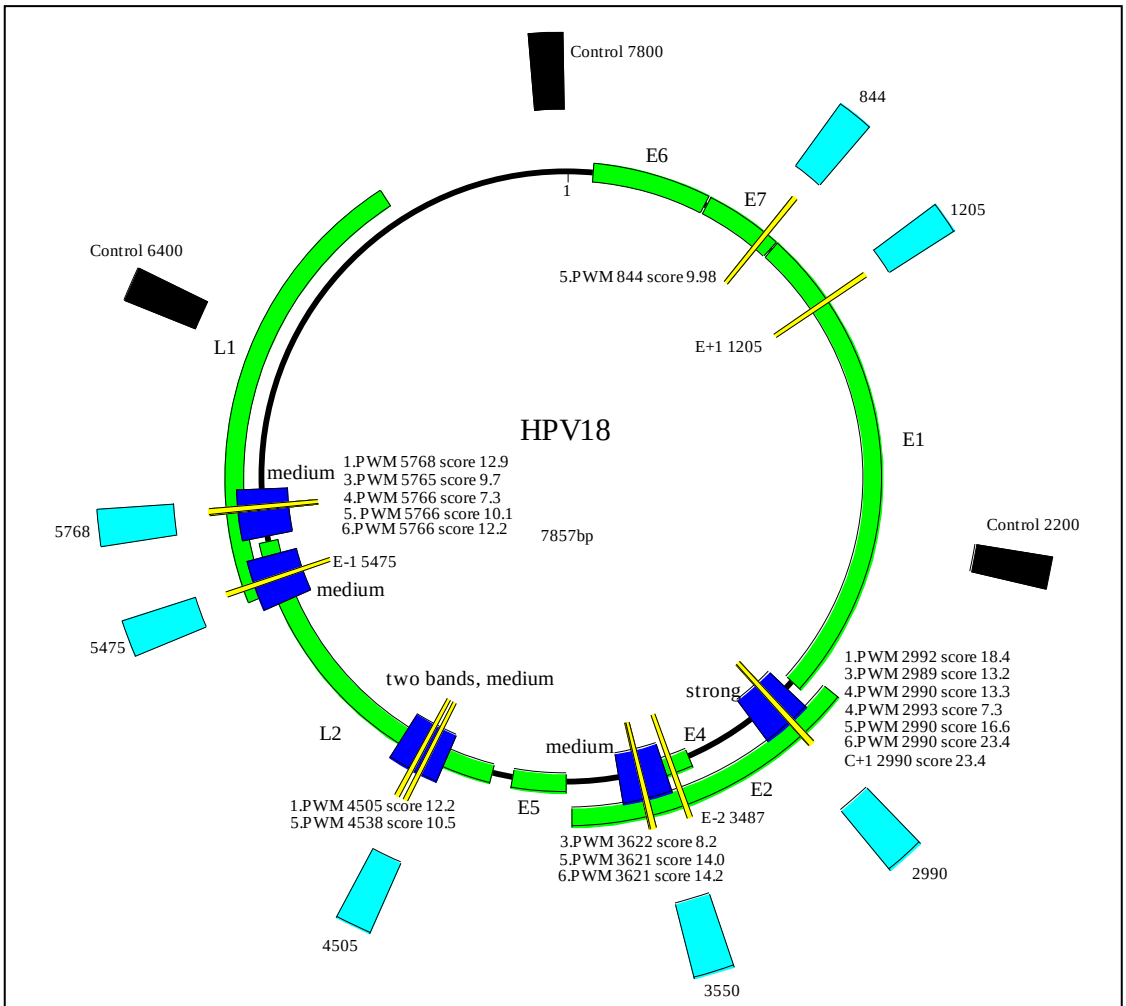
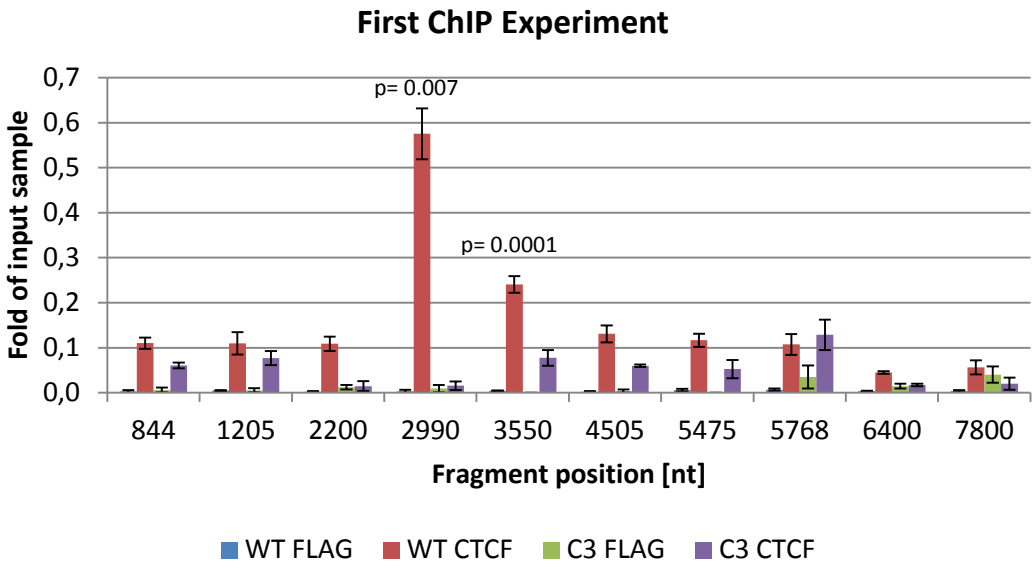


Figure 45) Fragments used for qPCR analysis of ChIP samples
The teal blocks show the fragments for testing predicted CTCF binding sites. Fragments indicated by black blocks have no CTCF binding site in their vicinity and were used as negative controls. Predicted CTCF binding sites are shown in yellow and fragments that showed a shift in EMSA are shown in blue.

A fragment of the c-Myc promoter that had been shown to bind CTCF was used as a positive control (Vetchinova *et al.*, 2006). The negative control was a fragment of the TLR2 gene that does not bind CTCF. Both of these fragments are amplified from host DNA. However, only the positive control gave a Ct value whereas the negative control remained below the Ct threshold after 45 PCR cycles. This was also experienced by a collaborator (Dr. Ian Groves from the Coleman laboratory, University of Cambridge, UK) who had independently encountered the same problem. Hence the fold enrichment of the positive control could not be determined. The normalisation of qPCR results was done against a 1:100 dilution of the input WT sample so that a local peak of a particular fragment was accounted as a confirmed CTCF binding site. Furthermore it was made sure that this local peak was not present in the ChIP sample precipitated with FLAG antibody.

The two ChIP experiments were performed on samples from undifferentiated Georgie cells and the results are shown in Figure 46 and the p values can be found in Table 36. The WT sample showed increased precipitation of fragment 3000 with CTCF antibody. This increase was not seen with the C3 mutant or with FLAG antibody. However other binding sites did not show a peak in the ChIP experiments despite the fact that these fragments showed weak or medium shifts in EMSA. Only the fragments adjacent to nucleotide 3000 show slightly increased precipitation with CTCF antibody which is most likely an artefact due to their close proximity to the confirmed CTCF binding site. Additionally, our collaborators at the University of Cambridge were able to confirm CTCF binding in the region of nucleotide 3000 in HPV16 by using HPV16-maintaining, undifferentiated W12 cells for ChIP (personal, unpublished communication). However the ChIP experiments only captured the CTCF binding pattern of a particular stage of the HPV life cycle. The pattern may change when the life cycle progresses to a later stage and EMSA confirmed CTCF binding sites may be used at a later stage.

A)



B)

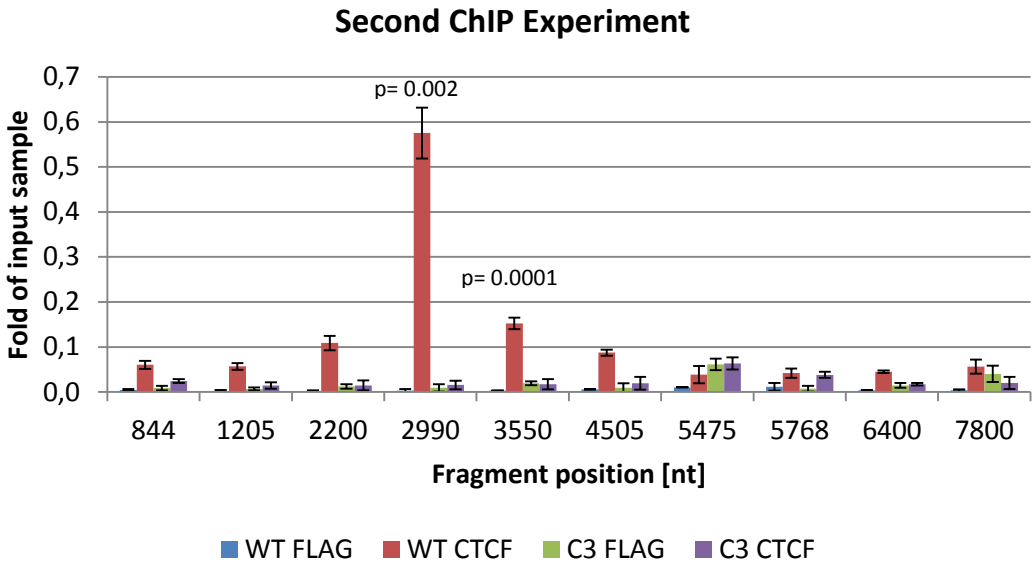


Figure 46) Results of two ChIP experiments
A sample of undifferentiated Georgie cells maintaining WT or C3 HPV18 was used for two ChIP experiments. The qPCR analysis of each ChIP experiment was done in triplicate. Precipitation with CTCF and FLAG antibody is shown. A peak is seen around nucleotide 3000 in the WT samples but not in the C3 mutant samples. All other fragments tested negative for CTCF binding even though some of them showed a medium or weak shift in EMSA. Only slightly increased precipitation with CTCF antibody is seen with fragments adjacent to nucleotide 3000 which is expected to be artefact caused by their close proximity to binding site 3000. All p-values can be found in Table 36. The samples analysed here are undifferentiated keratinocytes and the CTCF binding pattern may change upon differentiation. However reduced episomal copy number was seen in WT cells used for ChIP and this reduction was even stronger in C3 cells as shown in Figure 47. This experiment has been repeated by Dr. Jo Parish using cells of similar copy number and the results shown here were confirmed. Fragment 3000 in the WT binds CTCF and this binding is not seen in the C3 mutant.

		844	1205	2990	3550	4505	5475	5768
First ChIP experiment	GWT CTCF	0,911	0,924	0,007	0,000	0,301	0,776	0,814
	GC3 CTCF	0,414	0,397	0,828	0,417	0,311	0,353	0,073
Second ChIP experiment	GWT CTCF	0,443	0,278	0,002	0,000	0,149	0,085	0,043
	GC3 CTCF	0,119	0,585	0,799	0,996	0,826	0,017	0,009

Table 36) P-Values of the potential CTCF binding sites tested by ChIP
These p values were calculated using Student's double sided t-test against the null hypothesis that there is no difference in precipitation between the fragments tested and the 3 negative control fragments. In both experiments the fragments 2990 and 3550 of the WT show that this null hypothesis is not true for these two fragments. The spatial proximity of fragment 3550 to CTCF binding site 2990 is likely to have caused this outcome since the p value for this fragment is much higher in the C3 mutant. In the second ChIP experiment, the fragments 5475 and 5768 show evidence that the null hypothesis is false for these fragments. However fragment 5475 shows as much precipitation with CTCF antibody as with FLAG antibody. Thus it is likely that this is an experimental error. Fragment 5768 shows a very small standard deviation in the 3 measurements that skews the result into a low p-value.

In a later experiment the copy number of HPV18 episomes in the ChIP input samples was determined. By doing so it was discovered that a proportion of HPV18 episomes was lost in these samples. HPV episome loss is a known phenomenon that is not entirely understood (reviewed by Doorbar et al., 2012, Lace *et al.*, 2008, Stubenrauch *et al.*, 2000). However it is known to be an important step in cancer progression since the loss of episomal HPV genomes often abrogates the negative regulation over the viral oncogenes of HPV DNA that has randomly integrated into the host genome (Vinokurova *et al.*, 2008). The C3 episomes were lost more quickly than the WT episomes as shown in Figure 47. Approximately half of the wild type episomes were lost at the time when the ChIP experiments were carried out. The number of C3 episomes present in the ChIP samples was about 5 fold less than the number of WT genomes. This has to be considered when interpreting the ChIP results since the lower amount of C3 genomes could have influenced the result. However the binding of CTCF in the area around nucleotide 3000 and the abrogation of this binding in the C3 mutant in cells with high copy number has been confirmed in an independent ChIP experiment by Dr. Jo Parish (unpublished data). The loss of viral episomes observed in the ChIP sample gave rise to another set of experiment investigation this episome loss.

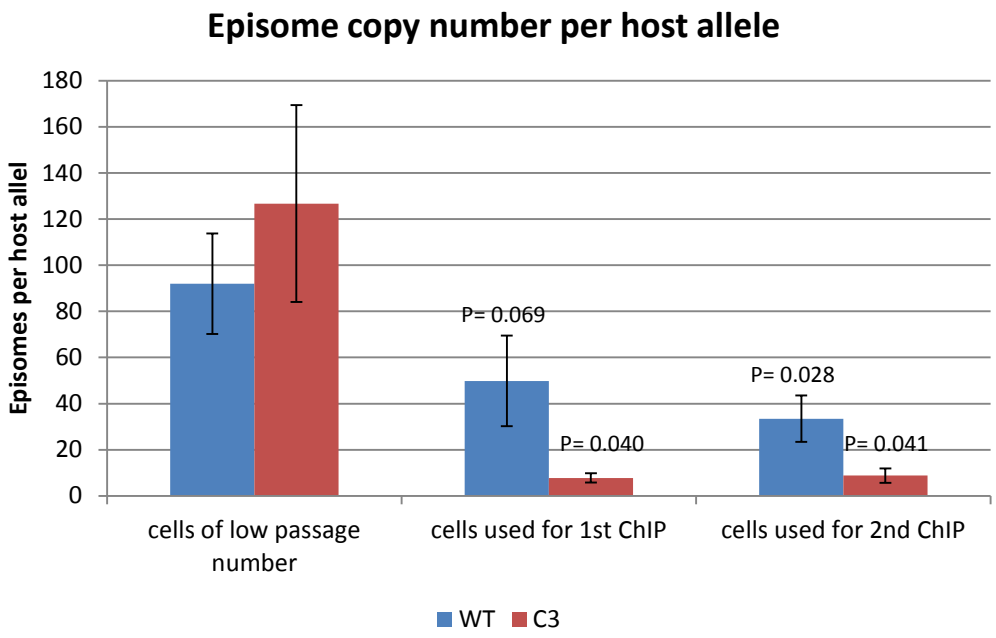


Figure 47) Episome copy number per host allele of cells used for ChIP
The episome copy number of every sample was determined in three repeats of the qPCR analysis. Both cell lines lost episomes over time. The WT had lost about 60 % of its episomes in the cells used for the ChIP experiments. The loss in the C3 mutant is about 90 %. The p values were determined using Student's double sided t-test against the null hypothesis that the episome copy number stayed the same since the first measurement at low passage number.

3.6.3 Determination of episome copy number in wild type and C3 mutant-containing HFK lines

Over the time course of an HPV infection it is known that the loss of viral episome can occur and the reasons for this are largely unknown (reviewed by Doorbar et al., 2012, Lace *et al.*, 2008, Stubenrauch *et al.*, 2000). In most cervical tumours HPV genomes have integrated into the host DNA followed by loss of episomally maintained HPV genomes (Yu *et al.*, 2005, Vinokurova *et al.*, 2008). In other viruses, like herpes virus Saimiri and Kaposi's sarcoma associated herpes virus, the abrogation of CTCF binding resulted in impaired maintenance of the viral genome (Stedman *et al.*, 2008, Zielke *et al.*, 2012). Thus episome copy number in the HPV infected cell lines was determined to check for potential episome loss due to a possible impairment in episomal maintenance caused by the lack of CTCF in the mutant cell lines. The copy number of a host locus and HPV18 episomes was quantified by qPCR using a sample of whole DNA extracted from the HPV18 maintaining cell lines with the QIAamp

DNA Mini Kit. The WT and C3 mutant cell lines from Georgie and Clonetics cells of various passage numbers were harvested with the intention to determine the episome copy number of each of these cell lines over a time course. Since the cells were split once a week, the passage number is also the number of weeks the cells were growing in culture. For the repeat experiment, cells of low passage number were de-frosted and cultured over an extended period of time. Thus, cell samples of all four cell lines from two different batches were available for analysis. The qPCR analysis of each sample was done in three separate reactions.

The episome copy number per host allele was determined mathematically by setting the Ct value of a fragment amplified from the viral episome in correlation to the Ct value of a fragment amplified from a single copy host locus. There are no large repetitive sequences in the HPV18 genome. Thus a primer pair targeting any location of the episome can be used for determining the number of episomes. The primers used for determining host locus copy number amplify a part of the TLR2 gene and have been previously used for episome copy number determination (Gray *et al.*, 2010). However, any single copy host gene can serve this purpose. Single copy genes are present as two alleles in each cell in G1, which needs to be considered when determining the episome number per host cell rather than per allele. The primer pairs chosen for episome copy number determination can be found in Table 37.

Template	Primer name	Location	Amplicon length	Sequence 5' to 3'	Tm
HPV18	HPV18 Episome FW	2167-2251	85	TTATAGGCGAGCCCAAAAAC	59,2
HPV18	HPV18 Episome RV	2167-2251	85	CCAATCTCCCCCTTCATCTAT	59,3
Human TLR2 locus	Human TLR2 FW	1893-1957	65	GCCAGCAAATTACCTGTGTGA	61.1
Human TLR2 locus	Human TLR2 RV	1893-1957	65	GGCGGACATCCTGAACCT	61.1

Table 37 Primers used for the determination of episome copy number
A fragment of the single copy locus TLR2 is targeted as well as a fragment of the HPV18 episome, accession number AY262282.1.

In a qPCR reaction these primer pairs were used together with whole DNA extracted from the HPV18 containing cell lines in combination with the SsoFast EvaGreen supermix from Bio-Rad. From the qPCR results a standard curve was generated for each primer pair and a common threshold for both primers pairs was chosen. Standard curves were checked for hallmarks of reliable qPCR like a slope close to -3.33 and a R^2 value close to 1. It was also ensured that the Ct value for every sample was within the range of the standard curve. Additionally the melting curve for each amplicon was checked to ensure that each primer amplified a single target. According to the common threshold chosen a Ct value for each of

the DNA samples could be determined for each of the primer pairs. The Ct value in combination with the formula of the standard curve contained enough information for mathematical determination of episome copy number.

Two different methods were used for qPCR-based episome quantification (Livak and Schmittgen, 2001, Pfaffl, 2001). These methods are called the Livak or $2^{-\Delta\Delta Ct}$ method and the Pfaffl method. However, both methods needed to be modified because they are typically used to determine the change in mRNA expression as a ratio of two samples, for example a treated and an untreated sample. Accordingly the ratio of mutant episomes per cell to wild type episomes per cell could be determined with the original formulas (Figure 48). In this case the ratio of 1 would mean that an equal amount of episomes is present in cells maintaining either wild type or mutant HPV18. Originally these methods were applied to qPCR results from cDNA samples. Both methods depend on the comparison of the target gene against a constant reference gene.

Livak Method:

$\Delta Ct_{(test)} = Ct_{(target\ of\ test)} - Ct_{(reference\ of\ test)}$

$\Delta Ct_{(calibrator)} = Ct_{(target\ of\ calibrator)} - Ct_{(reference\ of\ calibrator)}$

$\Delta\Delta Ct = \Delta Ct_{(test)} - \Delta Ct_{(calibrator)}$

$Ratio = 2^{-\Delta\Delta Ct}$

Pfaffl Method:

$$ratio = \frac{(E_{target})^{\Delta Ct_{target\ (calibrator-test)}}}{(E_{reference})^{\Delta Ct_{reference\ (calibrator-test)}}}$$

Figure 48) Original Livak and Pfaffl method
In the Livak method primer efficiency is assumed to be 2 whereas the Pfaffl method takes individual primer efficiencies into account.

In the case of episome quantification the original formulas of both methods can be used but some of the variables have to be re-defined. TLR2 takes the place of the reference gene. Instead of determining mRNA levels, the number of host genomes is determined using the single copy locus TLR2. The "target gene" in the modified method is not a single gene but the entire viral episome. Also two other components in the original methods need to be addressed. The "calibrator" is an untreated sample in which reference gene and target gene are expressed in an amount that will serve as a reference to the treated sample. In the modified method this calibrator is replaced by the cell sample that contains wild type HPV18 DNA. The second component of the original method is the "test sample" which is a treated sample in which gene expression is expected to change. The test sample is replaced by the cell sample maintaining C3

mutant HPV18. Thus the modified formulas give the ratio of C3 to WT episomes per host allele of TLR2. All these conversions are summarised in Table 38.

Component	Original methods for cDNA quantification	Episome quantification	Genes and samples chosen
Reference	cDNA of gene with constant expression levels	gDNA single copy host gene	TLR2
Target	cDNA of gene of interest	gDNA episome	HPV18 Episome
Calibrator	Untreated sample	Cell line containing wild type virus	GWT or CWT
Test	Treated sample	Cell line containing mutant virus	GC3 or CC3

Table 38) Factors used in the determination of either cDNA levels or episome copy number

The absolute number of episomes per host allele can also be determined by altering the modified formula of either method. It is simple to alter the Livak method because the exponent needed for calculating the number of either episomes or host alleles is already determined in the first steps of the original Livak calculation and just needs to be applied to the basis 2 to yield the absolute copy number of either viral episomes or host alleles. Altering the Pfaffl method is more complex due to the fact that the term for calculating the exponents needs to be transformed into fractions. Subsequently, the formula can be converted into a form where the numerator is the number of episomes and the denominator the number of TLR2 alleles. In G1 cells, two copies of the host locus are present whereas cells in G2 contain this allele four times. Thus episome copy number per cell is either two times or four times the result of either one of the altered formulas depending on the phase of the cell cycle.

The Livak and Pfaffl methods also differ in the way they incorporate primer efficiency. The primer efficiency is the factor by which the amount of amplicon is increased with every cycle. In theory, the optimal primer efficiency is 100 % (or 2 if given as a factor) which means that 100 % of the template present at the beginning of the cycle is amplified upon completion of the cycle. However, the primer efficiency is often not optimal. The acceptable range is between 90 % and 105 % (or factors 1.90 and 2.05). Primer efficiencies above 100 % are possible because the efficiency is calculated from the standard curve, which can be slightly inexact. Also nonlinear DNA amplification due to contamination can

influence the standard curve and therefore primer efficiency. The primer efficiency was calculated for each primer pair using the slope of the standard curve and the formula given in Figure 49. The Pfaffl method works with exact primer efficiencies whereas the Livak method assumes a primer efficiency of 100 % for every primer pair used. Thus the Livak method is easy to use but slightly inexact whereas the Pfaffl method is more exact but requires more complex calculations. Also the Pfaffl method depends on an exact standard curve. Small inaccuracies can easily cause a bias in the result.

Formula for primer efficiency:

$$\text{Primer efficiency} = 10^{\frac{-1}{slope}}$$

Example calculation: primer efficiency as factor:

$$10^{\frac{-1}{-3.298}} = 2.01$$

Example calculation: primer efficiency as percentage:

$$10^{\frac{-1}{-3.298}} - 1 = 1.01 = 10\ 1\ \%$$

Figure 49) Determination of primer efficiency

3.6.3.1 Using the Livak Method to compare wild type and C3 mutant episome copy numbers

Step 1: Normalisation of WT and mutant sample. The difference between the episome Ct value and the host allele Ct value is calculated.

$$\Delta Ct_{(WT)} = Ct_{(episome\ of\ WT)} - Ct_{(host\ gene\ of\ WT)}$$

$$\Delta Ct_{(mutant)} = Ct_{(episome\ of\ mutant)} - Ct_{(host\ gene\ of\ mutant)}$$

Step 2: The ΔCt value of the mutant sample is normalised against ΔCt of the WT sample.

$$\Delta\Delta Ct = \Delta Ct_{(mutant)} - \Delta Ct_{(WT)}$$

Step 3: calculating the ratio of C3 episomes per host allele relative to WT episomes per host allele.

$$Ratio = 2^{-\Delta\Delta Ct}$$

The complete modified Livak method for the determination of the ratio of C3 genomes to WT genomes with all previous steps combines is therefore:

$$2^{-\left((Ct_{episome\ C3} - Ct_{host\ gene\ C3}) - (Ct_{episome\ WT} - Ct_{host\ gene\ WT})\right)}$$

The ratio of C3 episomes to WT episomes was calculated using the Livak method. Samples of all cell lines from two different batches across different passage numbers were analysed and the results are shown in Figure 50. The data shows that this ratio declines over time. At low passage number the ratio was close to being equal but over the course of about 10 passages the ratio shifted towards a higher number of WT genomes than C3 genomes. This was observed in cells from two batches, each batch consisting of cells from two different donors. There are two possible reasons for this outcome. Either episome copy number in cells maintaining wild type episomes increased or episome copy number of C3 mutant episomes decreased. The absolute episome copy number was determined to find out which of the possible solutions is true.

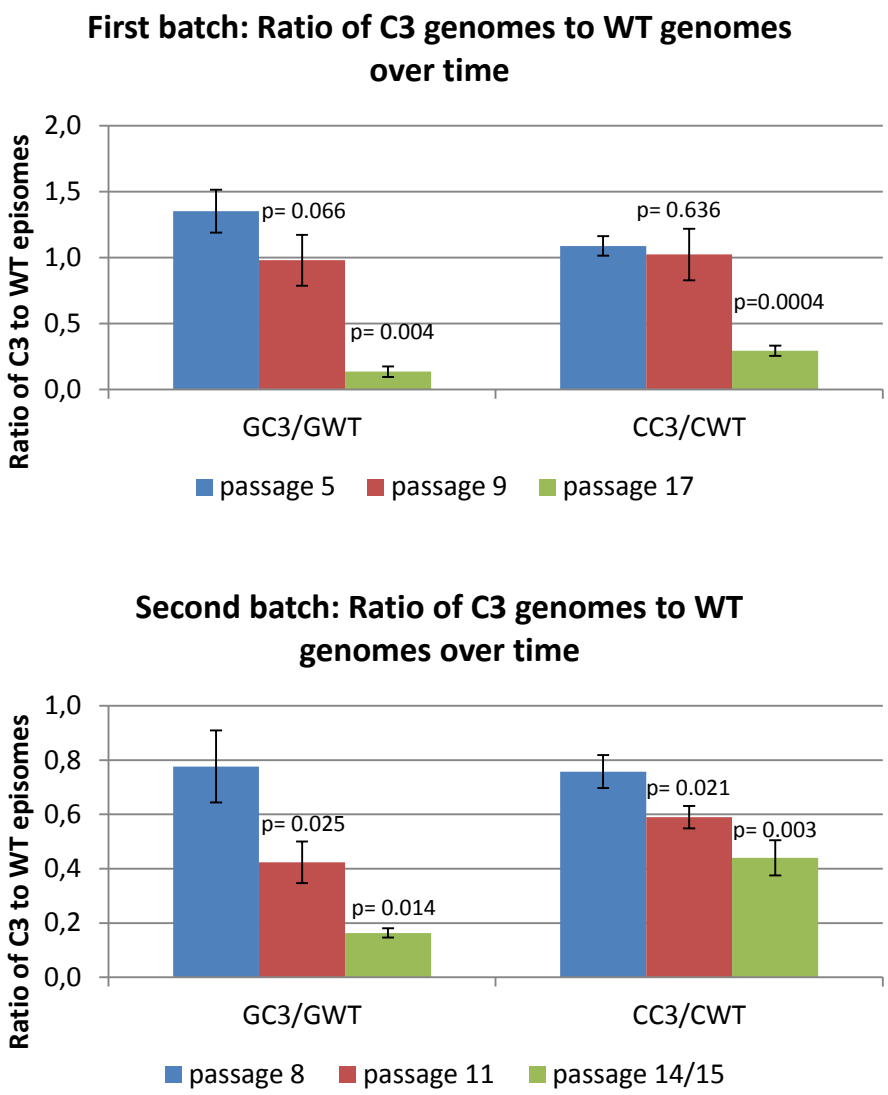


Figure 50) The ratio of C3 episome to WT episomes in HFK from both donors
Each batch is an independent experiment. The ratio of C3 to WT episomes was determined in three qPCR experiments. The p-values have been calculated using the two-sided Student's t-test with the null hypothesis that the ratio of C3 to WT episomes has remained the same as in the first measurement (passage number 5 and 8, respectively). A decline in the ratio of C3 episome to WT episomes in cells from both batches was seen. This decline was seen in samples from both donors but it is slightly faster in Georgie cells compared to Clonetics cells. The original Livak method was used to calculate the ratio of C3 mutant genomes per cell relative to wild type HPV18 containing HFK of the same passage number. Reasons for this outcome could either be an increase in WT episome or a decline in C3 episomes. To determine which of the possible solutions is true the absolute episome copy number was determined in a following experiment. These cell lines were split once a week so the passage number corresponds to the time of culturing in weeks.

Determination of the absolute episome number per host allele base on the Livak method:
The absolute number of episomes per host allele could be calculated using the previously determined ΔCt as the negative exponent to the primer efficiency, which is set to be 2 in the Livak method. Thus the formula of the absolute episome copy number per host allele is $2^{-\Delta Ct.}$ or

$$2^{-(Ct_{episome} - Ct_{host\ gene})}$$

The results of the episome copy number determination in cells of various copy numbers per host allele are presented in Figure 51. The episome copy number of cells of low passage number is slightly higher than the expected range of about 100 to 200 copies per cell (Frattini *et al.*, 1997). However a more recent review suggests the copy number to be more close to 200 copies (reviewed by Doorbar et al., 2012). All cell lines needed to be split about once a week so passage numbers and weeks of culturing are the same. The amount of viral episomes in the wild type cells of both batches showed little decline over the time course of 12 and 7 passages, respectively. On the other hand the cells maintaining the C3 mutant showed a much steeper decline in episome copy number over time. Only the GC3 cells from the first batch show a near-stagnation of viral episomes between passage 5 and passage 9. However the decline seen in the cells from the third time point is very steep. Only a small amount of C3 HPV18 DNA remained in GC3 cells of passage 17. A minimal amount of HPV DNA is thought to be essential for continuous cell growth due to the fact that immortalisation of the primary cells had been facilitated by the expression of the viral oncogene. If all viral DNA was lost only cells that acquired continuous cell growth through random mutations would be able to survive. However this viral DNA could be integrated rather than being episomal and that would prevent it from getting lost. The C3 mutant cell lines from both donor and both batches showed loss of episomes. However this loss is only statistically significant in Georgie cells. However the episome loss Clonetics cells containing C3 is close to being statistically significant with its p value of 0.053. Thus the reason for the shift in the ratio of C3 to WT episomes observed earlier is a faster decline of C3 mutant episome copies compared to the decline of WT episome copies. Donor specific traits may cause the different outcome in the Georgie and Clonetics cells containing the C3 mutant. The cell lines containing wild type HPV18 show a small decline in the average number of episomes, however this decline if far from being statistically significant.

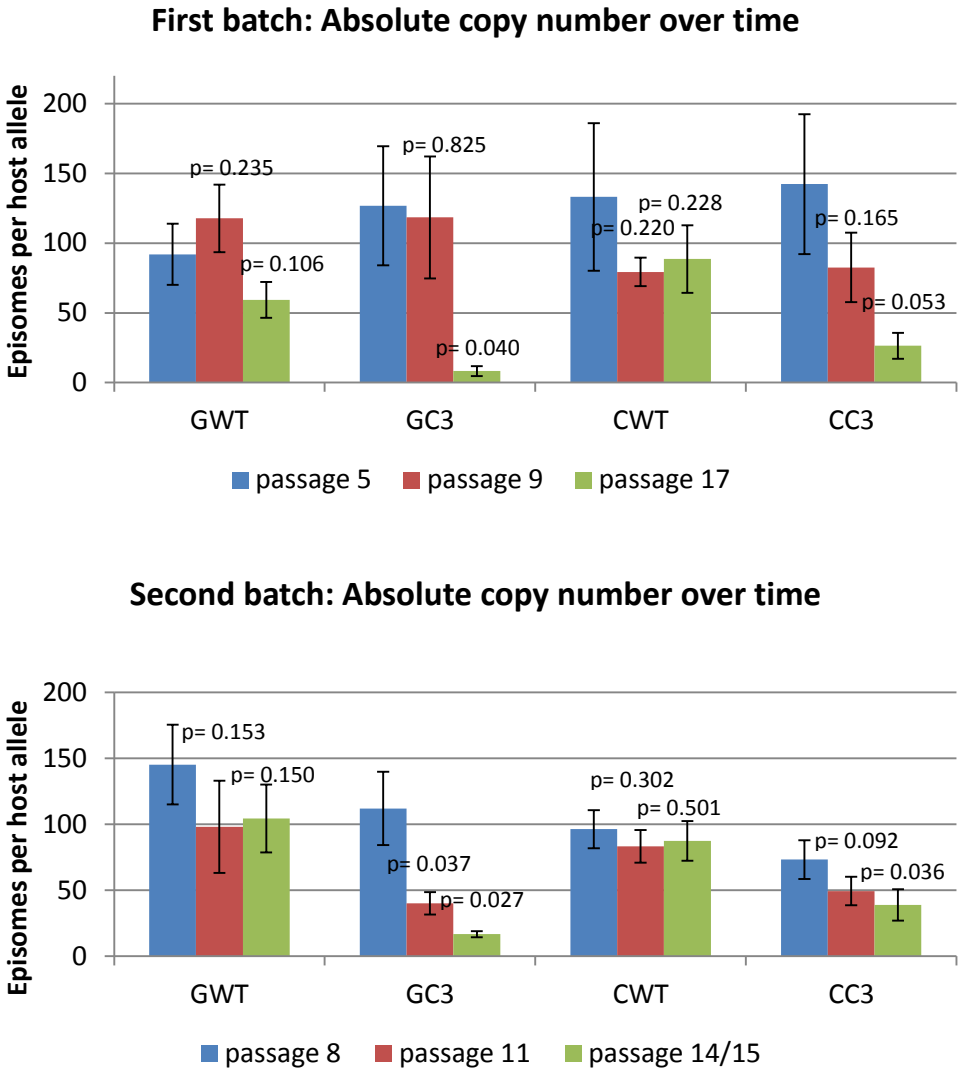


Figure 51) Determination of the absolute copy number of viral episomes
Each batch is an independent experiment. The episome copy number was determined in three qPCR repeats and the modified Livak method was used to determine the absolute copy number of viral episomes per host allele. The p-values were calculated using the two-sided Student's t-test with the null hypothesis that the number of episomes has remained the same as in the earliest passage number (passage 5 and 8, respectively). The pattern seen in both batches is similar. The number of WT and C3 episomes declines over time. However the decline in C3 genomes is stronger compared to WT genomes. This is seen in cells from both donors. Georgie cells loose C3 genomes slightly faster than Clonetics cells. Most WT samples show a stagnating number of episomes after about 10 passages. Only the GWT cells from the first patch show a further decline at passage 17.
These cell lines were split once a week so the passage number corresponds to the time of culturing in weeks.

3.6.3.2 Using the Pfaffl Method to compare wild type and C3 mutant episome copy numbers

The Pfaffl method works similarly to the Livak method but takes empirically determined primer efficiencies into account. Accordingly the Ct values of episome and host fragments are used as exponents to their particular primer efficiency. The fact that the primer efficiencies are likely to be different for both of these primer pairs makes the calculation more complicated compared to the Livak method because exponents to different bases cannot simply be subtracted from one another. The primer efficiencies for the host gene primer and the episome primer can be calculated by using the slope of the standard curve as shown in Figure 49. For the Pfaffl method the ΔCt values are calculated differently compared to the Livak method.

Step 1: Calculation of ΔCt values for host gene and episome.

$$\Delta Ct_{(\text{host gene})} = Ct_{(\text{host gene of WT})} - Ct_{(\text{host gene of mutant})}$$

$$\Delta Ct_{(\text{episome})} = Ct_{(\text{episome of WT})} - Ct_{(\text{episome of mutant})}$$

Step 2: Calculation of the ratio of C3 episomes to wild type episomes

$$ratio = \frac{(primer\ efficiency_{episome})^{\Delta Ct_{episome}}}{(primer\ efficiency_{host\ gene})^{\Delta Ct_{host\ gene}}}$$

The standard curves for the episome and TLR2 fragments have a slope of -3.314 and -3.330 respectively, both resulting in a primer efficiency of 2 which is the value that is used by default in the Livak method. Thus Livak and Pfaffl method yield the same result in this particular case. Hence the results of the Pfaffl method are not shown because they are virtually identical with the results from the Livak method shown earlier.

Determination of the absolute episome number per host allele based on the Pfaffl method:

Since the primer pairs targeting TLR2 and the HPV18 episome had different primer efficiencies, the Ct values cannot be simply subtracted from one another as is the case in the Livak method. Thus a more complex formula is needed. This formula can be derived from the original Pfaffl formula by converting the difference of the exponents into a fraction.

Original Pfaffl formula:

$$ratio = \frac{(E_{episome})^{\Delta Ct_{episome}}}{(E_{host\ gene})^{\Delta Ct_{host\ gene}}}$$

Conversion step 1: Substitution of ΔCt with the formula for ΔCt

$$ratio = \frac{(E_{episome})^{Ct_{WT\ episome} - Ct_{mutant\ episome}}}{(E_{host\ gene})^{Ct_{WT\ host\ gene} - Ct_{mutant\ host\ gene}}}$$

Conversion step 2: Converting exponents into fractions

$$ratio = \frac{\frac{(E_{episome})^{Ct_{WT\ episome}}}{(E_{episome})^{Ct_{mutant\ episome}}}}{\frac{(E_{host\ gene})^{Ct_{WT\ host\ gene}}}{(E_{host\ gene})^{Ct_{WT\ episome}}}}$$

Conversion step 3: The formula converted so that the fraction in the numerator defines the number of mutant episomes per host allele and the fraction in the denominator defines the number wild type episomes per host allele.

$$ratio = \frac{\frac{(E_{host\ gene})^{Ct_{mutant\ host\ gene}}}{(E_{episome})^{Ct_{mutant\ episome}}}}{\frac{(E_{host\ gene})^{Ct_{WT\ host\ gene}}}{(E_{episome})^{Ct_{WT\ episome}}}}$$

Thus the formula for the absolute episome copy number can be derived as follows:

Formula for calculating episome copy number:

$$\frac{(primer\ efficiency\ as\ factor_{(host\ gene)})^{CT_{(host\ gene)}}}{(primer\ efficiency\ as\ factor_{(episome)})^{CT_{(episome)}}} = \text{episomes per host allele}$$

Again, the fact that the primer efficiencies for both primer pairs were 2 led to virtually identical results of the Pfaffl method and the Livak method shown earlier. Hence these results are not shown again. However the modified Pfaffl method is still important if the episome copy number is determined using primer pairs for the episome and host locus that have different primer efficiencies. Primer efficiencies of each primer pair can also vary with changes in the qPCR setup like using a different master mix, primer concentration or temperature cycles.

3.6.4 Comparison of viral gene expression in wild type and C3 mutant by western blotting

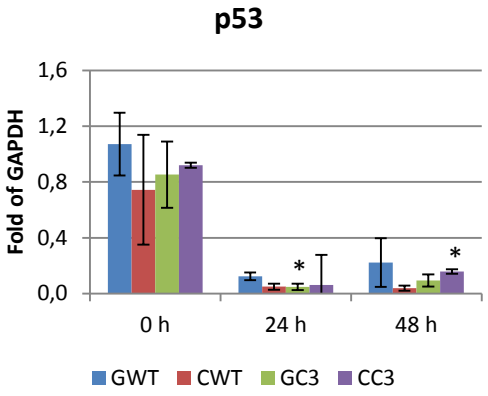
Human foreskin keratinocytes can be differentiated by growing them in methylcellulose (Wilson and Laimins, 2005, Green, 1977). Viral gene expression is coupled to host cell differentiation and analysing cells in different stages of the differentiation process gives the opportunity to compare viral gene expression over the course of the viral life cycle (reviewed by Doorbar, 2005). Thus the expression of viral genes can be monitored on methylcellulose differentiated cells using western blotting. Dr. Jo Parish provided cell samples of differentiated cells from each cell line previously generated. Those samples were made from cells of intermediate passage number (passage 11) from the second batch. Thus the copy numbers of WT and C3 episomes are relatively high as shown in the previous section. In three experimental repeats these cells were differentiated for 0, 24 or 48 hours. Terminal keratinocyte differentiation was achieved after 48 hours, so the viral life cycle should be completed within this time course (Green, 1977). However L1 expression of HPV18 in methylcellulose differentiated cells is inefficient and only occurs after extended culturing in methylcellulose (personal communication with Dr. Sally Roberts from the University of Birmingham). Western blots from the samples of each time point were prepared using antibodies targeting Glyceraldehyde 3-phosphate dehydrogenase (GAPDH), E2, E6, E7, E1^E4, loricrin, p53 and CTCF. Antibodies for E1 and E5 were either not available

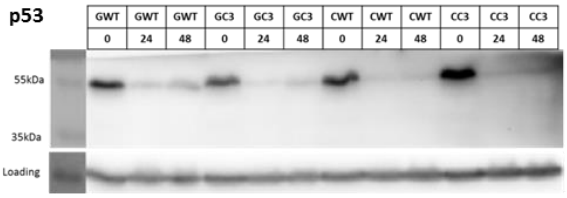
or not functional with the experimental conditions used. All bands were quantified and GAPDH was used as a loading control. Figure 52 shows the western blots performed to ensure appropriate differentiation of the cells and the western blot for CTCF. When keratinocytes start to differentiate the viral protein E6 efficiently targets p53 for degradation (Sherman *et al.*, 2002). This degradation of p53 was confirmed. However there were slight differences observed between WT and C3 (Figure 52). These differences are likely to be artefacts due to the background noise resulting from the quantification of very faint bands combined with the sometimes suboptimal quality of the western blot. Also the limited number of repeats influences the statistical significance of this result.

The host gene loricrin is a keratin that is expressed at the terminal stages of keratinocyte differentiation and was therefore used as marker to confirm the complete differentiation of the keratinocyte (Hohl *et al.*, 1991). A faint band of the expected size of 26 kDa in cells differentiated for 48h was seen (Gschwandtner *et al.*, 2013). The presence of loricrin confirms the terminal differentiation of the cell samples differentiated for 48 hours. Also the amount of CTCF in the cell throughout the differentiation process was analysed. All cell lines showed a constant amount of CTCF throughout the differentiation process. The band for CTCF became smeary in differentiated cell samples, which might be due to the accumulation of keratin in differentiated cells.

The western blots for E2, E1^E4, E6 and E7 are shown in Figure 53. Accumulation of E2 is expected at the intermediate and late stages of the differentiation process (reviewed by Doorbar et al., 2012). However the mutant cell lines show significantly increased E2 expression after 24 hours of differentiation compared to the wild type. Interestingly the mutated CTCF binding site is located within the E2 reading frame. The expression of E1^E4 is also altered in the mutant. E1^E4 is expected to be abundantly expressed at the late stages of the viral life cycle. Less E1^E4 expression is observed in the mutant cells lines after 48 hours compared to the cells containing the WT. However this reduction in E1^E4 expression is only significant in Georgie cells. In Clonetics cells the p value for the reduction of E1^E4 is 0.085. More experimental repeats are needed to double check the significance of the reduction in E1^E4 in Clonetics cells. It is possible that differential expression of E1^E4 in Georgie and Clonetics cells is a donor specific trait.

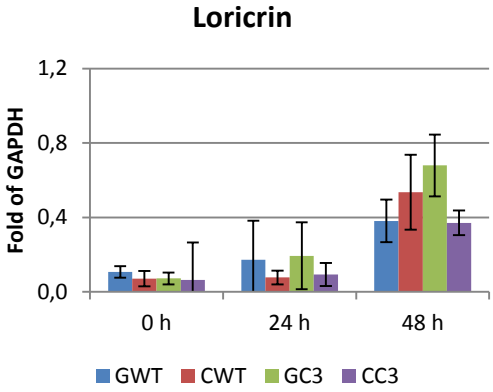
There is no difference in the expression of the viral oncogenes E6 and E7 between wild type and C3 mutant as shown in Figure 54. The band for E6 is strongest in cells differentiated for 24 hours. The expression of E7 is similar in all samples tested.

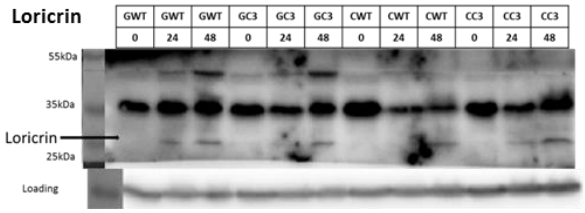




p-values

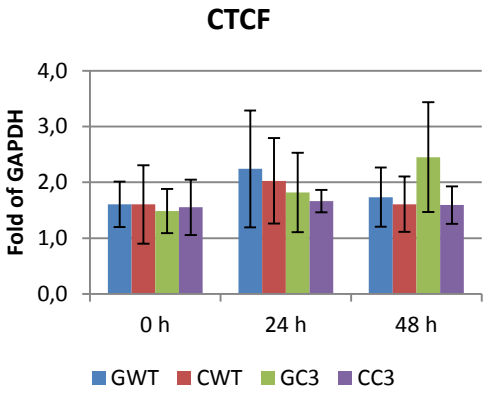
	Georgie C3			Clonetics C3		
	0 h	**24 h**	**48 h**	**0 h**	**24 h**	**48 h**
P53	0,313	0,022	0,327	0,543	0,486	0,035

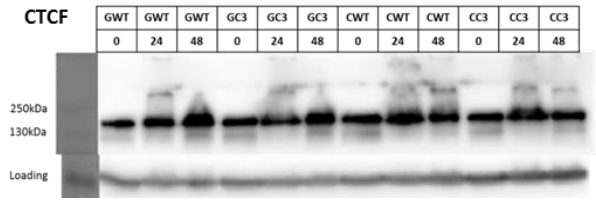




p-values

	Georgie C3			Clonetics C3		
	0 h	**24 h**	**48 h**	**0 h**	**24 h**	**48 h**
Loricrin	0,237	0,902	0,070	0,901	0,739	0,413

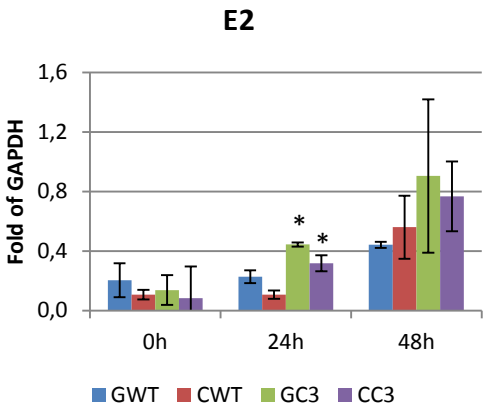


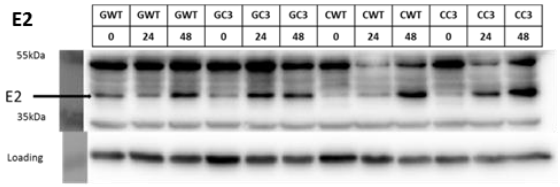


p-values

	Georgie C3			Clonetics C3		
	0 h	**24 h**	**48 h**	**0 h**	**24 h**	**48 h**
CTCF	0,733	0,599	0,346	0,910	0,510	0,963

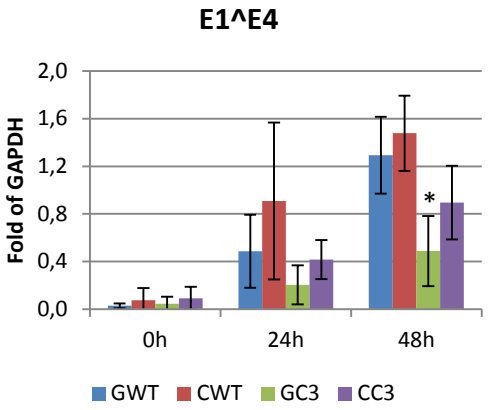
Figure 52) Differentiation controls and CTCF expression in Georgie and Clonetics cells
In three independent experiments, western blots of methylcellulose differentiated HFK were carried out using antibodies for p53, loricrin and CTCF. A representative western blot for each protein is shown. The bands from each of the three repeats for each protein were quantified and protein expression in WT and C3 cells was compared using the double-sided t-test against the null hypothesis, that the gene expression in WT and C3 is the same. p-values below 0.05 are indicated with a star in the graphs and with red background in the tables. The start of the differentiation process is confirmed by E6 mediated degradation of p53 after 24hours of differentiation. The C3 mutant shows slightly more efficient degradation of p53. However, this decrease is significant only for the 24 h sample of GC3 and the 48 h sample of CC3. Terminal differentiation is confirmed by the presence of loricrin after 48 hours. CTCF expression remains constant throughout the differentiation process in all samples tested.

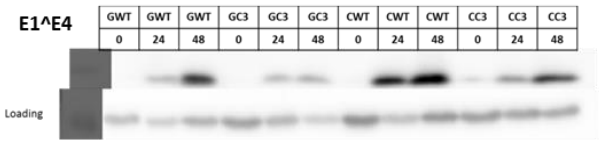




p-values

	Georgie C3			Clonetics C3		
	0 h	**24 h**	**48 h**	**0 h**	**24 h**	**48 h**
E2	0,497	0,008	0,260	0,652	0,000	0,372

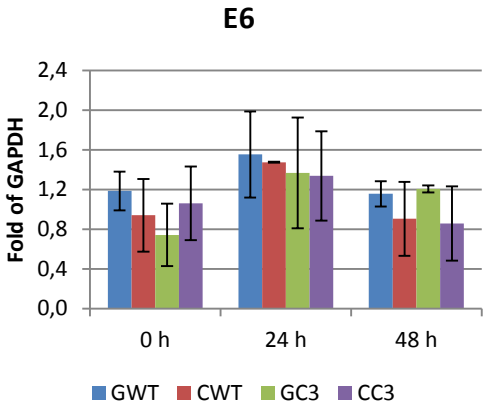




p-values

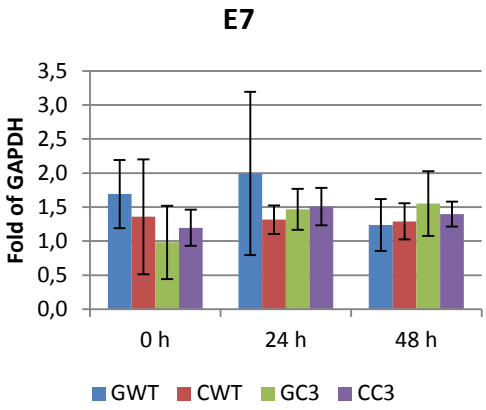
	Georgie C3			Clonetics C3		
	0 h	**24 h**	**48 h**	**0 h**	**24 h**	**48 h**
E1^E4	0,680	0,253	0,034	0,837	0,323	0,085

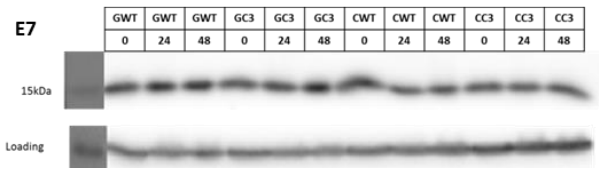
Figure 53) Gene expression of E2 and E1^E4 in Georgie and Clonetics cells
In three independent experiments, western blots of methylcellulose differentiated HFK were carried out using antibodies for E2 and E1^E4. A representative western blot for each protein is shown. The bands from each of the three repeats for each protein were quantified and the protein expression in WT and C3 cells was compared using the double-sided t-test against the null hypothesis that the gene expression in WT and C3 is the same. p-values below 0.05 are indicated with a star in the graphs and with red background in the tables. The expression of E2 is significantly increased in the C3 mutant after 24 hours of incubation and normalises again after 48 h. The expression of E1^E4 is significantly decreased in the GC3 sample differentiated for 48 h. In CC3, the same sample has a p-value of 0.85. A higher sample size is needed to reveal if the decrease in E1^E4 is also significant for CC3 cells.



p-values

	Georgie C3			Clonetics C3		
	0 h	**24 h**	**48 h**	**0 h**	**24 h**	**48 h**
E6	0,121	0,675	0,572	0,766	0,589	0,890





p-values

	Georgie C3			Clonetics C3		
	0 h	**24 h**	**48 h**	**0 h**	**24 h**	**48 h**
E7	0,171	0,530	0,424	0,776	0,295	0,682

Figure 54) Gene expression of E6 and E7 in Georgie and Clonetics cells
In three independent experiments, western blots of methylcellulose differentiated HFK were carried out using antibodies for E6 and E7. A representative western blot for each protein is shown. The bands were quantified and the protein expression in WT and C3 cells was compared using the double-sided t-test against the null hypothesis that the gene expression in WT and C3 is the same. p-values below 0.05 are indicated with a star in the diagrams and with red background in the tables. The expression of E6 peaks after 24 hours and all cell lines show the same expression pattern of this protein. The expression of E7 is the same in all cell lines across all stages of differentiation.

3.6.5 Monitoring organotypic raft culture morphology and gene expression using light microscopy and immunohistochemistry

Human foreskin keratinocytes can be differentiated to form *in vitro* generated skin in organotypic raft culture. This was done by growing the cells on a collagen plug with a limited supply of nutrients diffusing through this plug (Wilson and Laimins, 2005). Differentiating cells in this way enables the virus to go through its entire life cycle. In this study, raft cultures provided by Dr. Jo Parish were stained to compare cellular morphology and viral gene expression of wild type HPV18 maintaining keratinocytes and keratinocytes maintaining C3 mutant genomes. Two separate sets of rafts were grown from cells of the second batch with early passage numbers (below passage 8), so the episomal copy numbers of wild type and mutant are assumed to be approximately equal. The amplification of episomally maintained HPV18 genomes in these raft slides was confirmed using *in situ* hybridisation and showed no difference between wild type and C3 mutant (unpublished data from Dr. Jo Parish). The rafts grown from Clonetics cells were of higher quality than the ones grown from Georgie cells. Hence this section focuses mostly on the Clonetics rafts.
In HPV negative skin, only the basal cells actively divide. One of the daughter cells remains in the basal layer whereas the other daughter cell migrates towards the surface of the skin, differentiating on its way. The top layer of the skin consists of rigid, flat and keratinised cells which are eventually shed from the surface. Keratinocytes maintaining wild type HPV18 genomes also actively divide in the suprabasal layer in addition to the basal layer so the raft is generally slightly thicker compared to a raft grown from HPV negative cells (Blanton *et al.*, 1991). Figure 55 shows histological slices of raft cultures stained with eosin and hematoxylin. The raft grown from wild type maintaining cells showed the expected morphology (Frattini *et al.*, 1997, Lace *et al.*, 2009). However raft cultures grown from cells maintaining the C3 mutant showed even more cells in the basal, suprabasal and upper layer and the raft thickness was substantially increased in comparison to the wild type rafts. On the other hand the keratinous layer showed a decrease in thickness. Further investigation of this unusual raft morphology was done using immunohistochemistry to monitor viral gene expression in all layers of the raft. Prior to harvesting the rafts they were treated with bromodeoxyuridine (BrdU), which replaces thymidine during DNA replication and can be detected with a BrdU-specific antibody. Thus BrdU is used as a marker for cells that have actively replicated their DNA during the period BrdU was applied.

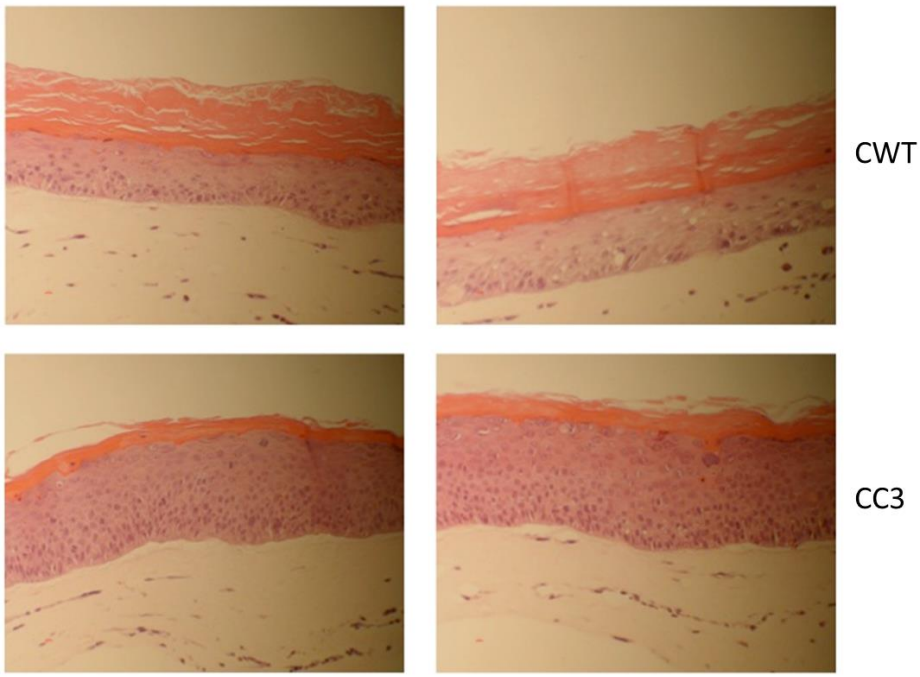


Figure 55) Histology of raft cultures grown from wild type and C3 maintaining HFK
Organotypic raft cultures were stained with hematoxylin and eosin. The images shown here are representative of two independent rafts for each of the two cell lines. The top pictures show rafts grown from wild type Clonetics cells. The rafts show typical morphology. Actively diving cells are found in a thin, basal layer and the suprabasal layer. Cells differentiate and flatten out as they migrate upwards until they form a dense keratinous layer. The bottom pictures show rafts grown from Clonetics cells with C3 mutant HPV18. Here the raft is substantially thicker. Differentiation and flattening of cells seems to be impaired and the keratinous layer on top of the raft is thinner compared to rafts grown with wild type maintaining cells. The rafts of the second donor (Georgie) showed the same morphological differences between WT and C3 mutant.

Fully grown raft cultures were embedded in paraffin before they were sliced with a microtome and mounted onto microscope slides. On every slide two primary antibodies were applied that were raised in different organisms so that two different secondary antibodies could be used to label each of the primary antibodies with a distinct fluorophore. In addition to the immunostaining, all slides were stained with Hoescht dye to visualise the nuclei by staining DNA. As a result each slide was stained for 3 targets of which each target was fluorescing at a distinct wavelength. In Figure 56 raft slides from all cell lines were stained with Hoescht stain and antibodies for BrdU and E1^E4. BrdU positive cells were seen in the suprabasal layer of cells maintaining either wild type or mutant HPV18. However the C3 mutant showed an increased number of BrdU positive cells in the upper layer.

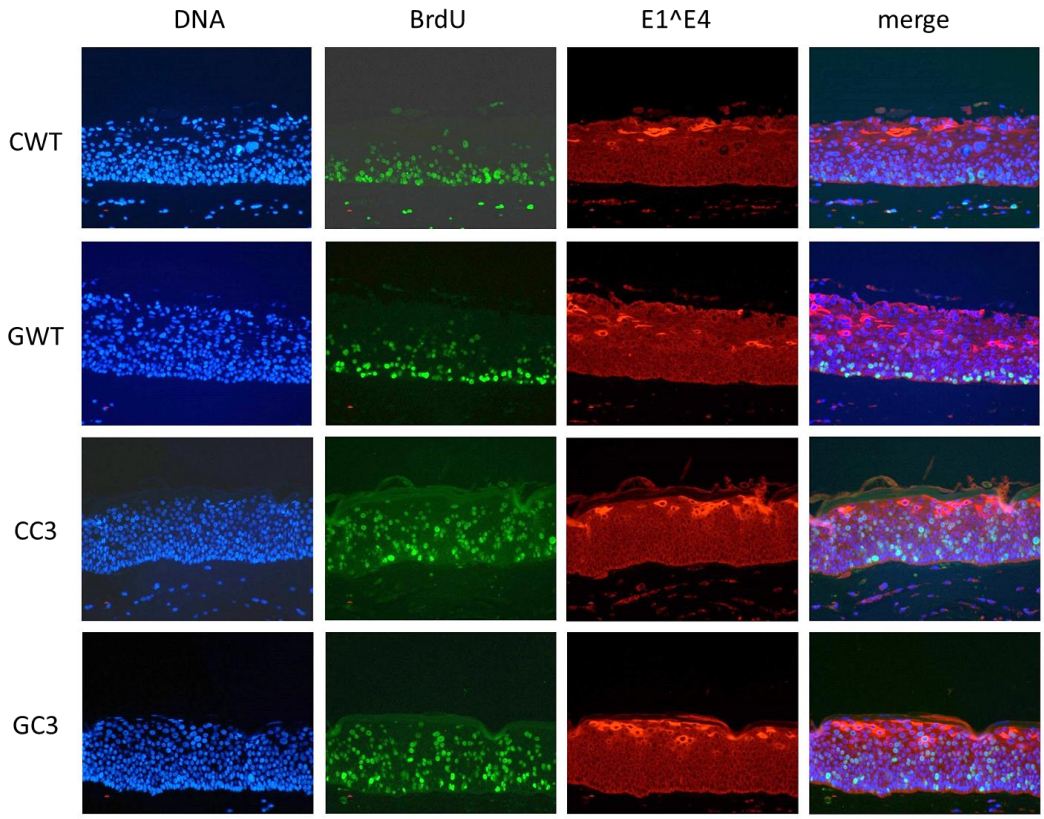


Figure 56) Raft slides stained for DNA, BrdU and E1^E4
Raft slides from all cell lines stained for DNA, the DNA replication marker BrdU and E1^E4. These images are representative of two independent experimental repeats. The wild type rafts show the expected pattern of BrdU positive cells in the basal and suprabsal layers. BrdU positive cells in the C3 mutant show a different pattern, they are found throughout the entire raft, including the upper layer. E1^E4 expression occurs in upper layers as expected and there is no difference in E1^E4 positive cells between wild type and C3 mutant.

The cell counts of BrdU positive cells showed that there is a significantly higher percentage of BrdU positive cells in the upper layer of the C3 raft as shown in Figure 57. The counts also revealed that the percentage of BrdU positive cells in the basal layer of the C3 is reduced compared to the wild type. Only the percentage of BrdU positive cells in the suprabasal layer is similar in WT and C3. To confirm the changes in DNA replication and supposedly mitosis, rafts were stained with antibodies for the mitosis marker Cyclin B1 and stained cells were counted (Figure 57 and Figure 58).

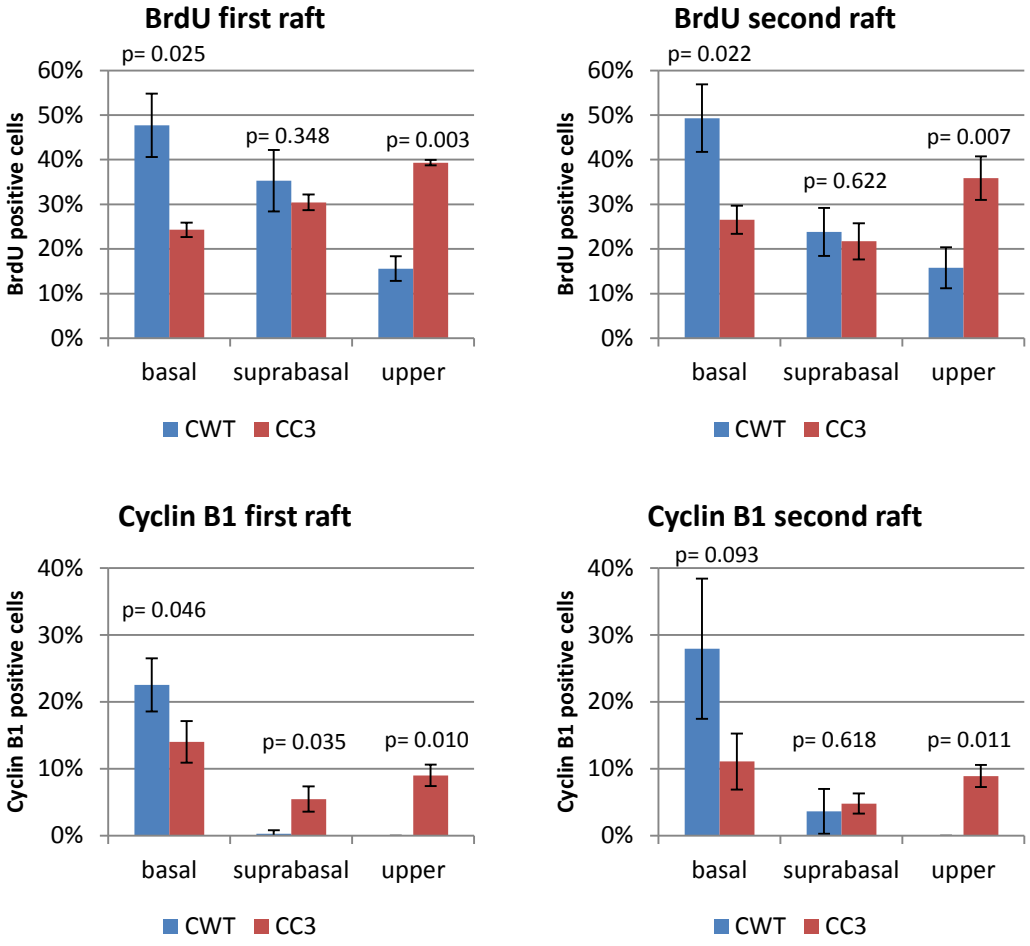


Figure 57) Quantification of BrdU and Cyclin B1 positive cells in CWT and CC3 rafts
Two independently generated organotypic rafts of each of the two Clonetics cell lines, CWT and CC3, were generated and stained using antibodies against BrdU and Cyclin B1. From each staining three representative pictures were taken and the percentage of stained cell was determined. The percentage of stained cells in CWT and CC3 was compared using the double sided t-test against the null hypothesis that the percentage of stained cells was the same in WT and C3. p-values are given on top of the bars.
There are significant differences in the staining of the DNA replication marker BrdU between WT and C3. In the basal layer of cells, significantly more DNA replication occurs in the WT. In suprabasal cells, the amount of DNA replication is about equal in WT and C3. In the upper layer of the raft significantly more DNA replication occurs in the mutant compared to the wild type.
Quantification of cells stained for the mitosis marker Cyclin B1 reveals a similar pattern. The basal layer of cells contains more cells undergoing mitosis in the WT compared to the mutant. However this difference is only significant in the first of the two repeat experiments (p=0.046 and p=0.93, respectively). Similarly, the quantification of Cyclin B1 positive cells in the suprabasal layer is also only significant in the first experiment. There are significantly more mitotic cells in the upper layer of the mutant rafts compared to the wild type rafts.

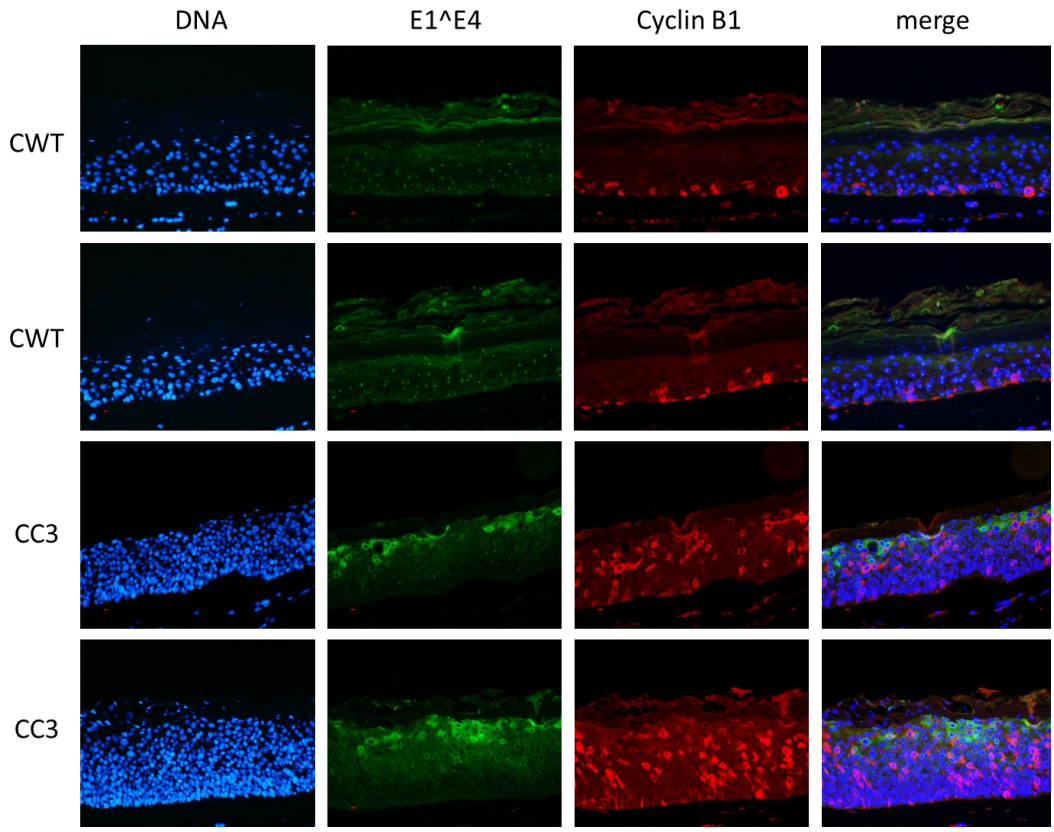


Figure 58) Raft slides stained for DNA, E1^E4 and the mitosis marker Cyclin B1
These images are representative of two independent experimental repeats. The C3 mutant slides show an increase in cell number. E1^E4 expression is very sporadic and counts of E1^E4 positive cells revealed no significant difference in E1^E4 positive cells between wild type and C3 mutant rafts Figure 60. Cyclin B1 expression in the wild type only occurs in the basal and suprabasal layer of the raft. In the mutant Cyclin B1 positive cells are found throughout the whole raft, including the upper layer.

The pattern observed in the BrdU staining was confirmed with the Cyclin B1 staining. A lower percentage of Cyclin B1 positive cells was found in the basal layer of the raft in the C3 mutant compared to the WT. However this difference was only significant in the first of the two repeat experiments (p=0.046 and p=0.93, respectively). Similarly, the quantification of Cyclin B1 positive cells in the suprabasal layer was also only significant in the first experiment (p=0.035 and p=0.618, respectively). More experimental repeats are needed to test if the difference in the percentage of mitotic cells observed in these layers consistently reaches statistical significantly. Considering the results of the BrdU staining and the p-values achieved in the Cyclin B1 staining, reduced proliferation of C3 positive cells in the basal layer seems likely. Regarding the Cyclin B1 staining in the upper layer, both repeat experiments agree that only the rafts containing the C3 mutant showed Cyclin B1 positively cells. Despite the reduced percentage of mitotic cells in one layer and an increase in

another layer, the overall thickness of C3 mutant rafts was increased. Thus the cells in all layers of the rafts were counted as shown in Figure 59. An increased number of cells was found in all layers of the C3 mutant rafts. Only the increase in the upper layer of one of the two experiments fails to reach significance with a p-value of 0.099. Combining this result with the staining for DNA replication and mitotic cells reveals that there are more cells in the basal layer in the C3 rafts, but fewer of these cells do still replicate their DNA and undergo mitosis compared to the WT raft.

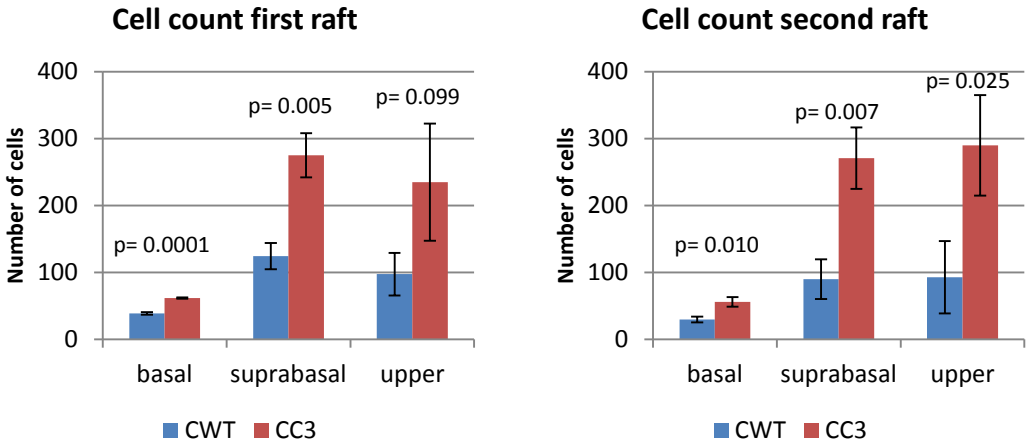


Figure 59) Quantification of cells in CWT and CC3 rafts
Two independently generated organotypic rafts of each of the two Clonetics cell lines were generated and stained using Hoescht stain to make the nuclei visible. From each staining three representative pictures were taken and the stained cells were counted. The number of cells in WT and C3 was compared using the double sided t-test against the null hypothesis that the number of cells was the same in WT and C3. p-values are given on top of the bars. Throughout all layers there are more cells in the C3 mutant raft than in the WT raft. Only the increase of cells seen in the upper layer of the first experiment did not reach significance (p= 0.099).

For the suprabasal layer, the percentage of replicating cells was shown to be equal in WT and C3 rafts while the overall number of cells in this layer was increased (Figure 57 and Figure 59). Regarding the upper layer, the increase in proliferation seen in the BrdU and Cyclin B1 staining agrees with the increased number of cells in this layer. To test if the C3 mutant generally accelerates the cell cycle, growth curves of undifferentiated cells in monolayer were generated by Dr. Jo Parish. These growth curves were very similar across all cell lines used (data not shown). Thus increased proliferation of cells infected with the C3 mutant only occurs in the differentiating cells of organotypic raft cultures.

The name E1^E4 suggests that this gene belongs to the early genes. However it is expressed in the upper and keratinous layers of the skin where the virus is in the intermediate and late stages of its life cycle (Wilson *et al.*, 2007). Staining for E1^E4 and quantification of E1^E4 positive cells revealed that those cells were seen in similar amounts in CWT and CC3 cell lines (Figure 60). The expression of this protein was rather sporadic and was not seen in every single cell. This is the typical expression pattern of E1^E4 (Griffin *et al.*, 2012). Even though western blots of methylcellulose differentiated cells showed a decrease in E1^E4 in CC3 cells, the rafts did not confirm this observation.

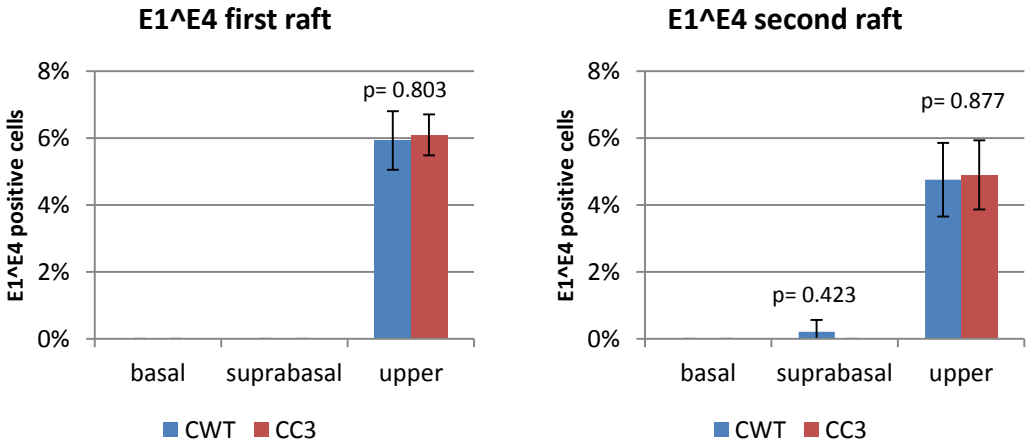


Figure 60) Quantification of E1^E4 positive cells in CWT and CC3 rafts
Two independently generated organotypic rafts of each of the two Clonetics cell lines were generated and stained using antibodies against E1^E4. From each staining three representative pictures were taken and the percentage of stained cells was quantified. The percentage of stained cells in WT and C3 was compared using the double sided t-test against the null hypothesis that the percentage of stained cells was the same in WT and C3. p-values are given on top of the bars. E1^E4 expression is about equal in WT and C3 and occurs in the upper layer of the raft.

Figure 61 shows raft slides stained for the keratin loricrin and E1^E4. The wild type showed fully flattened and differentiated keratinous cells that form a tight layer on top on the raft as it naturally occurs in skin (Nagarajan *et al.*, 2010, Barcelos and Sotto, 2009). Cells maintaining the C3 mutant however did not flatten out and the keratinous layer was spotty instead of tight. This combined with the increase of cell proliferation in C3 infected rafts implies that the morphological differentiation process of the host cell was impaired. The amount of cells stained for loricrin was not quantified since the cells maintaining the WT were fused together.

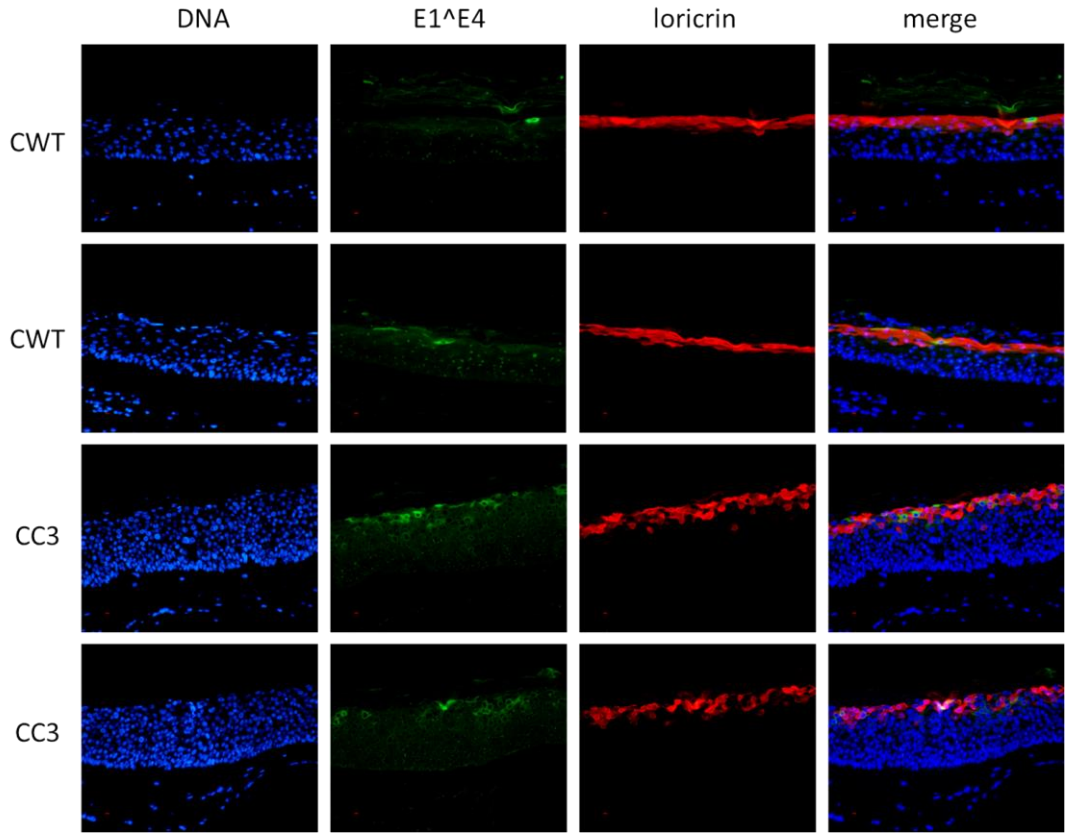


Figure 61) Raft slides stained for DNA, E1^E4 and the keratin loricrin
These images are representative of two independent experimental repeats
The C3 mutant slides show a significant increase in cell number. Loricrin expression occurs in the uppermost layer of the raft. In the wild type rafts, loricrin positive cells are flat and form a rigid surface on top of the raft. However loricrin positive cells maintaining the C3 mutant do not form this tight layer. Instead, loricrin positive cells have a round shape and the keratinous layer has gaps.

4 Discussion and Implications

4.1 *In silico* screening and optimisation of future predictions

The results presented in this study reveal that all three prediction tools used can provide a solid foundation for CTCF research. However the manual predictions using *Storm* gave the best results and the statistical analysis indicated how to optimise *in silico* prediction of CTCF binding sites further.

The Essex tool is based on EMSA confirmed CTCF binding sites. Therefore, the influence of host factors or DNA methylation is not taken into account in predictions with this tool. Thus, solely the potential of a CTCF binding site to bind CTCF without any cellular factors determines if a potential binding site is recognised by this tool. Approaching an *in silico* prediction like this has advantages as well as drawbacks compared to the other tools. For example, 30 % of CTCF binding site are suggested to not contain the 20mer consensus binding motif of CTCF and can therefore not be recognised by PWM based tools (Chen *et al.*, 2012a). However, an EMSA based tool does not neglect these binding sites. In many cases the CTCF binding pattern was shown to be subject to change upon cellular differentiation or imprinting (Bell and Felsenfeld, 2000, Hark *et al.*, 2000). Therefore, an EMSA based tool can find binding sites that may not be currently occupied by CTCF but have the potential of being used in response to a stimulus or during differentiation of the cell. In contrast to this, ChIP based tools are limited towards the CTCF binding pattern of the particular cells used to generate the ChIP data. A potential change in CTCF binding pattern that occurs in cells from other tissues, or in response to a stimulus, is not taken into account by ChIP based tools. On the other hand the ChIP tools take the cellular factors that influence CTCF binding in the particular cell type into account. These factors may be more or less abundant in different cell types and could therefore influence the CTCF binding pattern. This is an advantage of the ChIP based tools CTCFBSDB and *Storm* over the Essex tool. Using different tools based on EMSA and ChIP in parallel gives the opportunity to benefit from the advantages of each tool and compensate their flaws to some extent. Thus using EMSA and ChIP based tools for each new CTCF prediction is recommended.

By taking the statistical analysis into account it is possible to optimise the interpretation of results given by all three tools. Out of all the tools used, the Essex tool had the lowest ratio

of confirmed binding sites to predicted CTCF binding sites. However this hit ratio of 42.4 % was still higher than the hit ratio achieved with fragments that do not contain a predicted binding site (12.0 %). The CTCFBSDB tool 1.0 has a high hit ratio (64.5 %) but also a high chance of neglecting important binding motifs. The patched version 2.0 of the CTCFBSDB tool is an improvement over 1.0 but still neglects many important sites because only the best match to each PWM is shown. Unless this issue is resolved, predictions using *Storm* are the most reliable ones.

The identification of many binding sites combined with a high hit ratio (63.6 %) are a clear advantage of *Storm* over the other two tools because those either find a large number of binding sites (Essex) or have a high hit ratio (CTCFBSDB). Statistical analysis of the results from *Storm* can be used to improve future predictions. Out of the six different PWM, PWM 1 and 3 had nearly all their predictions above threshold confirmed *in vitro*, therefore, it would be sensible to decrease the threshold for these PWM in future experiments. For example the binding site below threshold found in the LCR of BPV1 was identified by PWM1 and has a score of 9.35 while the score threshold is 10.72. Based on the high hit ratio of motifs predicted by this PWM (92.9 %), a binding site with a score of 9.35 still has a high probability of being EMSA confirmed. On the other hand, an increase of the threshold for PWM number 5 would be sensible since this PWM had the comparably low hit ratio of 64.7 % at the threshold chosen.

However the best indicator for the EMSA confirmation of a predicted binding site is the number of different PWM by which this binding site was identified. All 13 fragments containing binding motifs that were recognized by 3 or more different PWM were shown to bind CTCF *in vitro*. In addition, the strength of the shift correlated with the number of PWM that identified a particular binding site and the score it was given by each PWM. The strongest shifts were mostly seen with binding sites receiving high scores from many PWM.

I recommend using *Storm* in combination with the Essex tool for new CTCF binding site predictions. However the CTCFBSDB tool could be on par with *Storm* if it is updated so that thresholds can be chosen manually and the neglect of binding sites is resolved. In this case the CTCFBSDB tool would give the same results as *Storm* since it uses the same PWM. Additionally it is widely available and easy to use. If the developers should decide to improve their tool this way it would benefit future CTCF research greatly.

However this study has already led to significant improvements of the CTCFBSDB tool by fixing the line break bug and the bug of giving incorrect location of putative CTCF binding

sites. The fixed tool now identifies the strong and conserved binding site around nucleotide 3000 in HPV31, which was neglected prior to the bug fix. It is thought that this particular motif had been disturbed by a line break and could only be discovered after the bug had been fixed. Since the CTCFBSDB tool is one of the most widely used tools for CTCF binding site prediction, fixing the bug will positively affect future research on CTCF.

Nevertheless using the prediction tools does not supersede experimental confirmation. Some binding sites found by EMSA, including the strong binding site for BPV1 at nucleotide 3000, were not predicted by any of the tools used. Considering the fact that all high risk HPV types have a confirmed strong, conserved binding site at this position makes it likely that the binding site in BPV1 in this position is also bound by CTCF *in vivo.* However more research is needed to confirm this.

4.2 The CTCF binding pattern of HPV

EMSA testing of predicted CTCF binding sites led to the generation of CTCF binding maps for each papillomavirus tested. The comparison of these binding maps revealed particular conserved CTCF binding sites. All high risk HPV screened have a high scoring CTCF binding motif around nucleotide 3000 that shows a strong shift in EMSA. Interestingly, no binding motifs were predicted in this area in low risk or beta HPV and EMSA testing for unpredicted binding sites in this region confirmed the absence of a CTCF binding site at this position in the low risk and beta types tested. However, both low risk HPV types tested had another high scoring motif predicted around nucleotide 5400 which showed a strong shift in EMSA. Most of the other viruses screened also had CTCF binding sites predicted in the wider area around nucleotide 5400 but the scores of these motifs were lower. EMSA testing of these binding sites showed shifts of either weak or medium strength. This, together with the lesser positional conservation of this binding site in high risk types, indicates that the CTCF binding sites around nucleotide 5400 may be more important to low risk types than to high risk types. The binding site around 5400 in low risk types may, in fact, replace the conserved binding site around nucleotide 3000 of high risk HPV types.

It has been found that CTCF binding sites of high occupancy are conserved in most types of tissue whereas sites of low occupancy are more likely to be tissue specific (Essien *et al.*, 2009). Assuming the strength of a gel shift is a measure of occupancy, the strong binding sites may be used throughout the differentiation process of the host cell whereas the sites

showing a weaker shift may be bound by CTCF in a particular stage of the differentiation process. Furthermore It has been shown that the degree of occupancy of a transcription factor binding site can have a specific function in gene regulation (Tanay, 2006). For CTCF there may be a distinct functional difference between binding sites of low and high occupancy (Essien *et al.*, 2009). The high occupancy binding sites of CTCF were shown to be more often associated with repressive histone marks whereas low occupancy sites were associated with transcriptionally active chromatin (Essien *et al.*, 2009).

The ChIP data presented in this study confirmed that the site in HPV18 at nucleotide 2990, which resulted in the strongest shift *in vitro*, bound CTCF in undifferentiated HFK. All CTCF binding sites that resulted in a weaker shift in EMSA were not confirmed by ChIP using undifferentiated HFK cells. Again assuming that shift strength is a measure of occupancy, this data supports the hypothesis made by Essien *et al.* that binding sites of high affinity are occupied regardless of tissue or stage of differentiation (Essien *et al.*, 2009).

The EMSA data produced in this study does not fully resemble *in vivo* conditions. CTCF is present in the gel shift reactions at high concentration and the binding reaction takes place in a binding buffer which is obviously different from *in vivo* conditions. Furthermore, the lack of co-factors and post-translational modifications in the *in vitro* binding reactions may play a role. The diversity in different binding motifs provided by the 11 zinc fingers could make CTCF inherently sticky to DNA which in turn may have led to some false positive results. I suggest that these false positive sites are more likely to show a weak shift in EMSA since the affinity to sub-optimal binding motifs should be low. Thus the strength of a shift may also be an indicator of how likely it is that an EMSA confirmed CTCF binding site actually binds CTCF *in vivo*; however this is speculation. Also the results achieved using ChIP have their limitations. EMSA confirmed binding sites with a medium or weak shift could potentially be below the detection limit of this technique and that could be another explanation as to why those weaker sites were not ChIP-confirmed.

Nevertheless, all EMSA data of high risk and low risk viruses combined led to the creation of a common pattern of CTCF binding for the viruses tested (Figure 62).

The area from nucleotide 7000 to 2400 was found to be completely devoid of CTCF binding sites in all HPV types tested. All of the binding sites predicted in this region were shown to be negative for CTCF binding in EMSA experiments. This is especially interesting since this region contains the whole LCR.

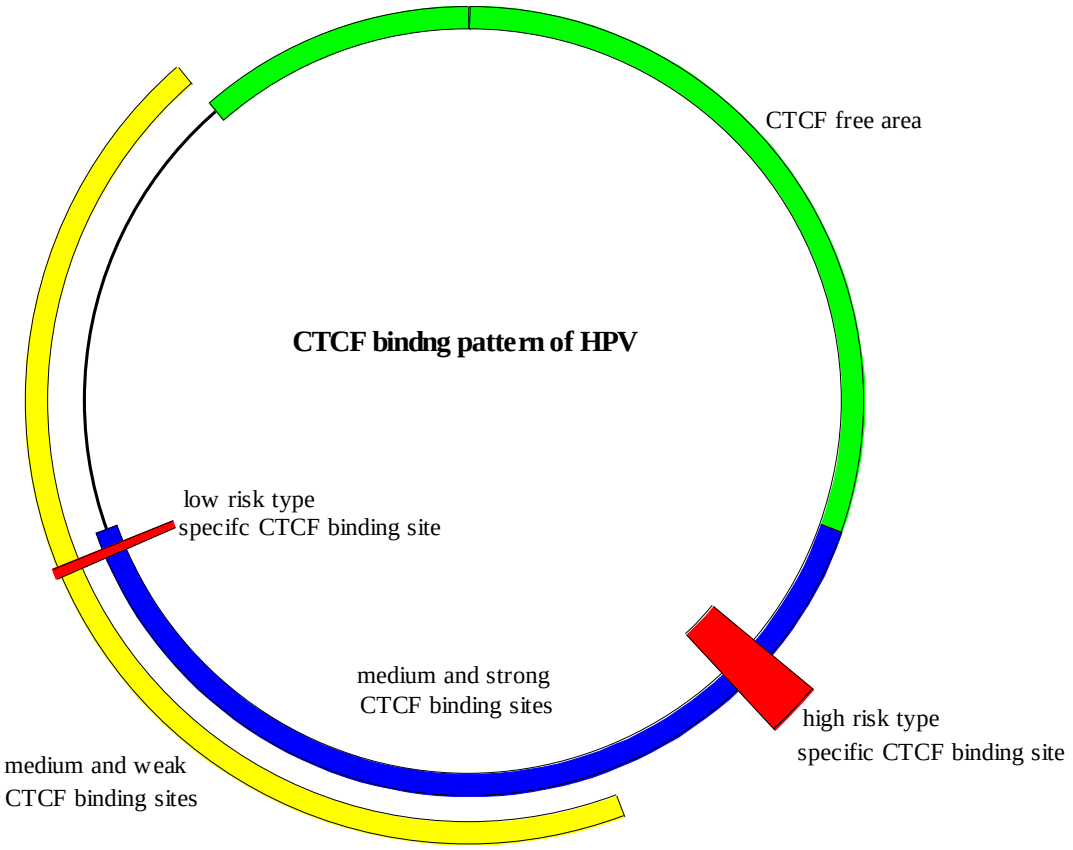


Figure 62) CTCF binding pattern of high and low risk HPV types
Out of 4 high risk HPV types and 2 low risk HPV types, none had a binding site confirmed within the area shown in green (nucleotide 7000 to 2400). CTCF binding sites showing a strong shift have exclusively been found within the blue area which is located from nucleotide 2400 to 5500. The red bar on the right marks the area where all high risk HPV have a conserved CTCF binding site that showed a strong band shift. Accordingly the red area on the left marks the area where both low risk viruses tested have a conserved and strong CTCF binding site. In the yellow area (nucleotide 3500 to 7000) CTCF binding sites showing a weak or medium shift are common. Many of these weaker sites accumulate here. However ChIP testing of HPV maintaining, undifferentiated keratinocytes did not confirm CTCF binding in this region.

However many functions of CTCF are long range so CTCF binding sites influencing the LCR may not necessarily be located within in the LCR (Zlatanova and Caiafa, 2009a). The sequence from nucleotide 2400 to 5500 was the only part of the genome where strong CTCF binding sites were found. Interestingly, this area spans the border between the early and the late genes so CTCF could potentially be involved in the separation of active from inactive chromatin. Also, CTCF binding sites that are highly conserved in high risk and low risk viruses are located in this region. CTCF binds strongly in the region of nucleotide 3000 in high risk viruses or nucleotide 5400 in low risk viruses. In EMSA, fragments containing these binding sites all showed a strong shift and the binding site around nucleotide 3000

was ChIP confirmed in undifferentiated cells maintaining HPV18 (Figure 46) and HPV16 (unpublished data from Dr. Ian Groves, University of Cambridge). The potential function of this binding site is discussed in detail in the following sections.

From nucleotide 3500 to 7000, the CTCF binding sites showed weaker shifts but binding motifs were abundant in this region and clustered in particular in the ORFs of the late genes. However ChIP experiments of undifferentiated cells maintaining HPV18 (Figure 46) or HPV16 (Dr. Ian Groves, University of Cambridge, personal communication) did not show CTCF binding to these sites. Thus CTCF may bind to these binding sites later in the viral life cycle or they may be false positives.

4.3 The CTCF binding pattern of BPV1

BPV1 was the most thoroughly tested PV in this study. This virus has served as a model for HPV gene regulation for the past 5 decades and is still widely used in research (Boiron *et al.*, 1964, Black *et al.*, 1963).

The CTCF binding pattern of BPV1 has some similarities to that of HPV but distinct differences have also been found. The most notable similarity to high risk HPV is the presence of the high risk specific, strong CTCF binding site around nucleotide 3000. However this binding site was not predicted by any of the tools used. The Essex tool predicted a binding site very close to the EMSA confirmed binding site, but this region was found to be a false positive in EMSA since fragment containing only parts of the predicted motif bound CTCF. Therefore, the true binding site seems to be located in the vicinity of the predicted site. The strong shift and the conserved location of this binding site make it likely that the EMSA result is not an artefact.

The remaining CTCF binding pattern within the BPV genome is distinct from that of all the HPV types tested. For example there are binding sites located within the area from nucleotide 7000 to 2400 and even in the LCR whereas this region appears devoid of CTCF binding sites in HPV.

Another interesting trait of BPV1 is the CTCF binding in the vicinity of promoters outside of the LCR. Each promoter outside of the LCR was shown to have a CTCF binding site in close proximity, most often located shortly upstream of the promoter judging from the predictions and the coverage of fragment P890. In KSHV it was shown that CTCF is necessary for recruiting RNA polymerase II to a specific promoter (Kang *et al.*, 2013). The

CTCF binding sites in BPV1 may have the same function. Also other functions of CTCF could influence the transcription of BPV1 genes, for example by blocking enhancers, forming loops or preventing *de novo* methylation.

The strong CTCF binding site around nucleotide 3000 may have a distinct role regarding the expression of different variants of the E2 protein. In BPV1, three E2 variants exist: Full length E2, E2-TR and E8^E2 (Hubbert *et al.*, 1988). The isoforms are expressed in a ratio of 1:10:3, respectively (Hubbert *et al.*, 1988). Full length E2 can activate or inhibit transcription from promoters in a dose dependent manner (Dostatni *et al.*, 1991, Steger and Corbach, 1997, Thierry and Howley, 1991). E2-TR is transcribed from its own promoter at 3080 and mostly functions as a repressor through competition with full length E2 for binding site (Lim *et al.*, 1998, Vaillancourt *et al.*, 1990). This variant of E2 is lacking the transactivator domain; however it was recently suggested to be able to still activate some promoters upstream of its binding site (McBride *et al.*, 1989, Lace *et al.*, 2012). Elimination of E2-TR resulted in an increase in viral copy number as well as increased transformation potential (Riese *et al.*, 1990, Lambert *et al.*, 1990). The expression of E2-TR is unique to BPV and the strong CTCF binding site is located between nucleotide 2917 and 3027 which is very close to the E2-TR specific promoter at nucleotide 3080 (Vaillancourt *et al.*, 1990). Thus this promoter may be regulated by CTCF. If this is the case CTCF would regulate transcription of the most prevalent variant of the viral transcription factor E2, indicating that CTCF binding in this region may be vital for the global regulation of viral gene transcription. However the western blots show that this is not the case for methylcellulose differentiated HFKs. A downregulation of E2-TR should result in increased viral copy number and increased transformation potential whereas upregulation of E2-TR is likely to have the opposite effect and may even repress viral gene expression to an extent that is insufficient for viral maintenance (Riese *et al.*, 1990, Lambert *et al.*, 1990).

The E2 variant E8^E2 is a clear repressor of E2 function similar to the ones found in HPV. The transcript encoding this variant is transcribed from promoter P890 (Vaillancourt *et al.*, 1990, Lace *et al.*, 2008). E8^E2 can repress transcription as well as replication of viral DNA (Lace *et al.*, 2008). CTCF may play a role in the regulation of E8^E2 expression considering that a CTCF binding site in the P890 promoter region was confirmed by EMSA.

Data from HPV18 and HPV11 showed that increased expression of E8^E2C resulted in maintenance defects by over-abundance of heterodimers of E8^E2C and E2 which are not sufficient for long term maintenance of the viral episome (Kurg *et al.*, 2010). However,

disruption of the E8^E2 protein in BPV did only show little effects on viral gene expression so the mechanisms from HPV16 and HPV31 may not apply to BPV1 (Lambert *et al.*, 1990).
Another feature of the CTCF binding pattern in BPV1 is the distribution of CTCF binding sites across the entire BPV1 genome. Considering that CTCF was found to protect an area of at least 2kbp around its binding site from *de novo* methylation, the distribution of CTCF binding sites in BPV1 has the potential to protect the majority BPV1 genome from *de novo* methylation (Zampieri *et al.*, 2012, Davalos-Salas *et al.*, 2011, reviewed by Klenova and Ohlsson, 2005, Fedoriw *et al.*, 2004, Szabo *et al.*, 2004). On the other hand the potential of CTCF to form DNA loops and bridges opens up the possibility that CTCF may be involved in packaging of the DNA into the viral particle by folding the episome in a particular way.
BPV1 was the only papillomavirus of which the whole LCR was tested for CTCF binding experimentally. The fragments used for screening the BPV1 LCR mostly had an overlap of 60bp but a few had a shorter overlap of about 30bp. Considering that the overlapping fragments for testing the binding site at nucleotide 3000 revealed that a CTCF binding motif can be very large, it is possible that a CTCF binding sites have been missed out due too short overlaps.

4.4 Generation of HFK cell lines maintaining an HPV18 mutant deficient in CTCF binding

The generation of a HPV18 mutant that was deficient in CTCF binding around nucleotide 3000 was successful. A near complete abrogation of CTCF was confirmed in EMSA experiments. Subsequent ChIP experiments suggested the abrogation of CTCF binding in C3 mutant maintaining HFK. However the low episome copy number of the C3 samples used for the ChIP could have influenced this result. Hence the ChIP experiments shown in this study have been repeated by Dr. Jo Parish with cells of normal episome copy number and the loss of CTCF binding in the C3 has been confirmed.
It is important to note that the fragment used in EMSA to test the binding site around nucleotide 3000 was 200 bp in size and may have contained a CTCF binding site other than the predicted one as it was the case for the fragments used for mutating binding site 5400 or testing binding site 3000 in BPV1. The fact that CTCF binding was abrogated in the mutant confirmed that the predicted binding site within the 200bp fragment tested by EMSA was the true binding site of CTCF. However, the confirmed binding site at nucleotide

3000 in HPV18 was also present on the 60bp probe, which did not bind CTCF in EMSA experiments. This could be explained by the motif being at the very edge of the probe what prevented CTCF binding to the motif. This is in agreement with the conclusion of Nakahashi *et al.* who suggested that some zinc fingers of CTCF bind nonspecific sequences at either side of the binding motif (Nakahashi *et al.*, 2013). Thus the absence of such a nonspecific sequence at one edge of the probe rendered CTCF unable to bind to the motif.

The mutation introduced at binding site 5400 of HPV18 failed to abrogate CTCF binding. This binding site was predicted by the Essex tool which had the lowest hit-ratio of all tools used and the fit to the PWM defined by Schmidt *et al.* was not optimal. Therefore it is possible that this binding site was a false positive prediction and that the true CTCF binding site exists elsewhere on the 200bp fragment used for testing the binding site around nucleotide 5400. However the ChIP data showed that this binding site is not occupied by CTCF in undifferentiated keratinocytes. Another incorrect prediction of a CTCF binding site was revealed in BPV1 for the binding site around nucleotide 3000. A putative binding site in the close vicinity of the real binding site was predicted by Essex tool. The actual location of the real CTCF binding site was shown to be not exactly at the predicted motif. The reasons for this could be mere chance or an inaccuracy of the Essex tool itself. However more data is needed to investigate this further.

Nevertheless, it was shown experimentally that unpredicted CTCF binding motifs exist on the fragments used for testing binding site 5400 in HPV18 and binding site 3000 in BPV1. With the thresholds used for *Storm*, no putative CTCF binding sites on these fragments were revealed. A new prediction with *Storm* could be performed to screen for binding sites using lower thresholds than the ones previously chosen. Considering the fact that binding sites above threshold predicted with PWM 1 and 3 were almost always confirmed, even binding sites below threshold recognised by these PWM have a good chance of revealing the real binding site on these fragments.

In this project only the mutation for CTCF binding abrogation around nucleotide 3000 in HPV18 was successful and could be used for further experiments. However three potential binding sites of other transcription factors at the mutation site were changed when the mutations were introduced as seen in Table 34. The binding site for MyoD was disrupted with the introduction of the mutations. MyoD is not expressed in HFK, so an effect on viral gene expression due to the disruption of this binding site is unlikely (unpublished data from Professor Ciaran Woodman, University of Birmingham). The change in the affinity of the

binding site for Oct1 is minor, resulting in a score increase from 0.964 to 0.982. However an effect on the viral gene expression cannot be excluded. A new potential binding site for GATA3 was created at the mutation site. This transcription factor is expressed in HFK and could therefore influence future experiments (unpublished data from Professor Ciaran Woodman, University of Birmingham). To investigate this further a ChIP experiment could be carried out to determine if GATA3 binding to the newly generated potential binding site occurs.

Creating a continuous, viral episome maintaining cell line from primary keratinocytes was a very labour intensive process which was done by Dr. Jo Parish from the University of Birmingham. Primary foreskin keratinocytes from two donors (Georgie and Clonetics) were chosen to reduce bias from donor specific genetics and to keep the workload to a manageable amount. Using primary cells of two donors has been shown to be beneficial for this project since a difference between the cell lines from different donors have been observed (Figure 50). The maintenance of the C3 mutant episome appeared to be worse in Georgie cells compared to Clonetics cells.

Before the creation of the cell lines it was not known if the mutated CTCF binding site at nucleotide 2990 of HPV18 performs a vital function to the virus. The successful creation of the C3 mutant cell line proves for the first time that the CTCF binding site around nucleotide 3000 in HPV18 is not vital for episome maintenance in the first weeks of culturing. However long term maintenance was affected as cells maintaining the mutant lost viral episome quicker compared to cells maintaining wild type HPV18. Sequencing of the mutation site from DNA isolated from all wild type and mutant cell lines confirmed the desired sequence in each of the samples. Thus there was no carryover of wild type HPV18 genomes in the C3 mutant sample.

Southern blot analysis of DNA from the cell lines which was digested with the single cutter *Eco*RI showed one clear band of the expected size for the linearised full length HPV18 genome. No background bands were seen. The samples were also digested with *Bgl*II and analysed by Southern blotting. This enzyme does not have a restriction site within the HPV18 genome and resulted in the typical banding pattern of circular DNA in open circle and supercoiled form on a Southern blot. A few very faint bands were seen in the background in addition to the two expected bands. If these bands were due to integrants such bands would also be seen in the *Eco*RI sample since *Eco*RI also cuts the host genome. However the Southern blot with *Eco*RI showed a clear background. Additionally the pattern

of these background bands was identical in all cell lines. Every integration event would take place at semi-random sites resulting in a different banding pattern after digestion (Thorland *et al.*, 2003, Dall *et al.*, 2008). Hence the faint bands seen in the *Bgl*II digests were considered to be background noise. Thus the presence of episomal HPV18 genomes in all cell lines was confirmed.

4.5 The role of CTCF in genome maintenance

The C3 mutant maintaining HFK have been shown to lose episomes over time. In other studies episome loss has been observed in response to various different alterations in HPV gene expression and the putative causes of episome loss can give indications to the function of the CTCF binding site around nucleotide 3000 in HPV18. The regulation of the episome copy number in HPV maintaining cells is complex and not fully understood. Viral gene expression, methylation, tethering and host cell traits contribute to this mechanism (Hoffmann et al., 2006, Parish et al., 2006, Kurg et al., 2010, Penrose and McBride, 2000, De-Castro Arce et al., 2012).

At the early passages 5 and 6 the samples of C3 Georgie and C3 Clonectics cells have been shown to have a similar copy number to their wild type counterparts. These cell lines needed to be split once a week so the cells of passage numbers 5 and 6 were already growing for over a month without significant episome loss. However C3 maintaining cell lines above 10 passages showed a decline in episome copy number (Figure 51).

The replication of the viral episome can be divided into three phases and the regulation of viral copy number is likely to be different in each of these phases (reviewed by McBride, 2008). Firstly the initial amplification upon infection of the cell, secondly the maintenance replication that keeps viral copy number in basal cells constant and thirdly the amplification of viral genome for the assembly of viral particles (reviewed by McBride, 2008). There is also some evidence for HPV latency at low episomal copy number which may involve a fourth phase of replication but HPV latency is not well understood at the current state of research (reviewed by Doorbar, 2013). The low copy number that remained in the C3 cells after extended culture could be an indicator for such a latency phase. However the immune system is suggested to be an important factor in HPV latency and this factor was not present when genome loss occurred in this study (reviewed by Doorbar, 2013).

The two major regulators of episome copy number in HPV are E1 and E2. Additionally the viral oncogenes E6 and E7 contribute indirectly to viral genome replication by the promotion of proliferation and the degradation of inhibitory factors such as TBP, CDP, YY1 and p53 (Hartley and Alexander, 2002, Ilves *et al.*, 2003, Narahari *et al.*, 2006). The regulation of episome copy number immediately after infection and later in the life cycle for the assembly of viral particles is dependent on E1 since this protein is suggested to facilitate episome replication independently of host DNA (Egawa *et al.*, 2012, Bodily and Laimins, 2011, reviewed by Doorbar et al., 2012). E1 is recruited to the origin of replication by E2, also making E2 essential for the replication of viral genomes in these two phases (Sedman and Stenlund, 1995). However the maintenance phase, which is in-between the initial amplification and the mass production of episomes, has been shown to be independent of E1 expression since replication takes place together with the host chromosomes (Egawa *et al.*, 2012, Kim and Lambert, 2002). HPV episomes in undifferentiated keratinocytes are in this maintenance phase so E1 protein is likely not essential for replication of viral episomes. Even the E1-mediated initial amplification of the HPV18 genome after transfection of the HFK may not have been necessary since many genomes were possibly transfected into the cell at once. However E1 has been shown to impair the tethering mechanism by interacting with E2, possibly leading to episome loss (Voitenleitner and Botchan, 2002). Thus induction of E1 expression has the potential to decrease viral copy number by impairing tethering of the viral genome to host chromosomes. The expression of E1 in the cell lines has not been tested due to the lack of functional antibody. However a putative change in transcription of E1 that is directly caused by the lack of CTCF would have taken place from the time the virus was transfected into the host cell but the episome loss observed started months after initial transfection. Only an indirect effect of the abrogation of CTCF binding could cause a change in E1 expression after a long period of time, for example the gradual accumulation of DNA methylation.

In contrast to E1, the protein E2 is involved in all three phases of episome replication, including the maintenance replication of HPV18 which takes place in primary keratinocytes (reviewed by McBride, 2013). In the natural HPV life cycle, E2 is crucial for recruiting E1 to the origin of replication of the HPV episome in the initial and late amplification phases. E2 binding sites in the LCR are needed for the recruitment of E1 and the number of binding sites necessary for stable episome maintenance differs between PV types (Piirsoo *et al.*, 1996, Li *et al.*, 1989, reviewed by McBride, 2013). For HPV18 three out of the 4 binding sites

for E2 are essential for episome maintenance (unpublished data from van Doorslaer *et al.*; mentioned in a recent review (McBride, 2013)).

In addition to its function in replication, E2 also mediates the maintenance of a stable episome copy number by tethering HPV genomes to mitotic host chromosomes (You *et al.*, 2004). E2 binds to its binding motif within the HPV genome and interacts with other proteins such as Brd4 and ChlRI. This interaction is needed to form a stable connection with a host chromosome similar to the one in sister chromatid cohesin (Parish *et al.*, 2006, McPhillips *et al.*, 2006, Kurg *et al.*, 2006). Note that Brd4 is only essential for tethering in some papillomaviruses, including BPV1 and some beta papillomaviruses (McPhillips *et al.*, 2006). In all alpha papillomaviruses tested (HPV11, 16, 31, and 57), the interaction of Brd4 and E2 is only relevant for gene regulation whereas the tethering mechanism is independent of Brd4 (McPhillips *et al.*, 2006). Furthermore the tethering pattern seems to be different in alpha virus as the association of episomes with host chromosomes is only seen in telophase and the general physical connection is suggested to be weaker (McPhillips *et al.*, 2006). So there is some degree of episomal co-localisation with host chromosomes in alpha papillomaviruses that is dependent on E2. Thus a change in E2 expression could also impair tethering and therefore genome maintenance. Interestingly the amount of E2 was shown to correlate with episome copy number in BPV1 and an increase in E2 also increased the copy number of viral episomes (Penrose and McBride, 2000).

In methylcellulose differentiated cells it was discovered that E2 expression was slightly increased earlier in differentiation in the C3 mutant compared to the wild type (Figure 53). However the amount of E2 present in undifferentiated cells in which the episome loss occurred could not be measure with western blotting because E2 expression in these cells is below the level of sensitivity with the E2-specific antibody available. It is possible that the E2 expression level may have also been increased in these cells. However, increased expression of E2 should increase viral copy number instead of decreasing it.

Several E2 binding sites within the LCR have been shown to be critical for the replication and maintenance of viral episomes (Li et al., 1989, Piirsoo et al., 1996, reviewed by McBride, 2013). The E2 binding motif $ACCG(N)_4CGGT$ contains a CpG motif that can be methylated and methylation of E2 binding sites results in abrogation of E2 binding (Kim et al., 2003, Thain et al., 1996). Therefore methylation of E2 binding sites can be linked to episome replication and maintenance. Interestingly, the viral proteins E2, E6 and E7 have

been shown to induce *de novo* methylation of the viral episome (De-Castro Arce *et al.*, 2012). In turn, this resulted in reduced expression of the methylation-promoting proteins E2, E6 and E7 and a reduction in episome copy number (De-Castro Arce *et al.*, 2012). All three of these methylation-promoting proteins are essential to genome maintenance in HPV16 and 31 suggesting a mechanism in which self-regulation of viral copy number is due to an interplay of methylation-promoting viral proteins and the regulation of transcription and episome copy number by methylation (Park and Androphy, 2002, Thomas *et al.*, 1999).

The majority of this methylation dependent regulation of viral copy number originates from the LCR, which is located about 3000 nucleotides away from the CTCF binding site around nucleotide 3000. The methylation protection function of CTCF applies to at least 2kbp around its binding site so it could potentially extend to the LCR, especially if CTCF would form a loop to the LCR (Davalos-Salas *et al.*, 2011, Zampieri *et al.*, 2012, reviewed by Klenova and Ohlsson, 2005, Szabo *et al.*, 2004, Fedoriw *et al.*, 2004). However, methylation of the ORF of a gene can also result in epigenetic silencing. In conclusion a change in methylation caused by loss of CTCF binding could contribute to the episome loss observed.

Methylation has recently been shown to be involved in splicing. Thus *de novo* methylation as a result of the abrogation of CTCF binding could result in altered mRNA processing of viral transcripts, leading to a loss of episomes (Maunakea *et al.*, 2013). In addition, CTCF mediated stalling of RNA polymerase II was shown to promote the assembly of a splicing complex, favouring the inclusion of weak upstream exons (Shukla *et al.*, 2011). The absence of CTCF would abrogate RNA polymerase II stalling which could result in alternative splicing and may change viral gene expression in a way that does not properly support episome maintenance.

Considering the role of CTCF as a boundary factor, it is possible that CTCF at nucleotide 3000 prevents the spread from the transcriptionally silent heterochromatin of the late genes into the early genes during maintenance phase (reviewed by Ohlsson et al., 2010, Probst *et al.*, 2009). Downregulation of early genes via this mechanism could also cause episome loss.

All papillomaviruses express at least one variant of E2 that can act as a repressor of full length E2 by forming heterodimers (reviewed by McBride, 2013). In alpha HPV this repressor is called E8^E2C since it contains a small part of the otherwise untranslated E8 ORF attached to the DNA binding domain of E2 (Hubbert *et al.*, 1988). E8^E2C does not contain the transactivator domain of full length E2. Mutation of the E8 ORF to prevent

E8^E2 expression in HPV16 and HPV31 revealed that long-term episome maintenance of HPV31 failed after an initial increase in viral copy number of 30-40 fold in HFK and SSC13 (HPV-negative squamous cell carcinoma line) (Lace *et al.*, 2008, Stubenrauch *et al.*, 2000). Episome copy number of HPV16 in HFK was also increased when E8^E2 expression was prevented. However unlike the HPV31 genomes, long term maintenance of mutated HPV16 genomes was supported in these cells (Lace *et al.*, 2008). Interestingly heterodimers of HPV18 and HPV11 E2 and E8^E2C have been shown to be able to initiate replication but fail in the long term maintenance of the HPV18 and HPV11 genomes in SiHa (squamous cell carcinoma line) and HaCaT cells (cultured human keratinocytes) (Kurg *et al.*, 2010). Thus both the depletion of E8^E2C and the overexpression of E8^E2C can lead to episome loss in some HPV types. Hence a putative over or underexpression of E8^E2C is one possible explanation for the accelerated loss of episomes observed in the C3 mutant.

The different outcomes regarding E8^E2C-dependency of episome maintenance in different HPV types and cell lines may be connected to the mode of replication the virus acquires in these particular cell lines. HPV genomes in the maintenance phase can be replicated through two different modes (Hoffmann *et al.*, 2006). The first mode is random choice episomal replication which means that the viral genomes replicate randomly to a certain final number during S-Phase (Hoffmann *et al.*, 2006). The second mode is called licensed replication and refers to a mechanism in which each viral genome is replicated once during S phase, just like host chromosomes (Hoffmann *et al.*, 2006). The mode of replication that HPV acquires was shown to be dependent on HPV type and the cell line used. HPV16 genomes replicate once per cell cycle in W12 cells whereas HPV31 replicates via random choice in CIN612 cells (cervical intraepithelial neoplasm cell line) (Hoffmann *et al.*, 2006). However, transfection of HPV16 or HPV31 into the spontaneously immortalised normal keratinocyte cells (NIKS) promotes random choice replication in both HPV types (Hoffmann *et al.*, 2006). In addition to E8^E2 expression, the level of E1 expression may be the determining factor of the mode of replication because high expression of E1 has been shown to induce random choice replication in HPV16 maintaining W12 cells (Hoffmann *et al.*, 2006).

The mechanism of genome replication in HPV18 maintaining HFK has not yet been investigated. If replication in HFK is licensed, the virus may not be able to replace single genomes if they are lost. Hence a small chance of losing an episome during each cell

division would result in a slow decline in the number of viral episomes over time. The loss of CTCF binding may result in such a chance for genome loss during cell division.

The varying effect of E8^E2C loss regarding maintenance in different cell lines may be connected to the mode of replication the virus acquires in these cell lines. Hence E8^E2C may only be essential to only one mode of replication, for example licenced replication. Unfortunately the study investigating the influence of E8^E2C on replication and the study investigating the mode of replication used different cell lines (Hoffmann *et al.*, 2006, Lace *et al.*, 2008). The mode of replication of the cells in which E8^E2C was depleted was not determined. Hence it is not known if the ambiguous effect of E8^E2C on episome maintenance is linked to the mode of replication.

Interestingly, E8^E2C is transcribed from a particular promoter in HPV16, 18 and BPV1 situated around nucleotide 1150 (Lace *et al.*, 2008, Hubbert *et al.*, 1988, reviewed by Van Doorslaer et al., 2013). The regulation of this promoter has not yet been researched and CTCF has the potential to affect transcription from this promoter remotely by blocking enhancers, protection from methylation, or looping. Hence abrogation of CTCF binding could potentially alter transcription of E8^E2C which could be the reason for the impaired episome maintenance seen in the C3 cell lines.

The expression of some E2 transcripts is partly regulated by the use of suboptimal codons in the E2 coding region (Oliveira *et al.*, 2006, Kurg *et al.*, 2010). With the mutations introduced in the C3 mutant, three codons in the E2 ORF were changed which may have resulted in an optimisation of codons, thus increasing translation efficiency of E2. However E2 consists of more than 360 amino acids and changing only three of them probably has limited impact on translation efficiency. The mutations were situated in the transactivator domain of E2 which is excised in the splice variant encoding E8^E2C. Hence the codon optimisation from the mutation of the CTCF binding site would not affect translation efficiency of E8^E2C through the optimisation of codons. Also, increased expression of E2 in the C3 mutant cells was only seen at a particular time point, after 24 hours of differentiation.

Another putative cause for the aberrant maintenance of episomes in the CTCF binding defective HPV18 genomes can be hypothesised on the basis that CTCF has been shown to play a role in asynchronous replication of alleles (Bergstrom *et al.*, 2007). The data published by Bergstrom *et al.* shows that the CTCF binding allele of the H19/Igf2 locus replicates late in S phase whereas the methylated allele that does not bind CTCF replicates early in S phase (Bergstrom *et al.*, 2007). Assuming the HPV episomes follow the same

principle, the replication of the C3 mutant episomes would be shifted from late S phase to early S phase. Considering that genome amplification in suprabasal layers takes place in a pseudo S-phase in which the early proteins E6 and E7 promote proliferation while the E1^E4 protein inhibits proliferation, the shift of genome replication from a late to an early stage of the S phase may have an impact on vegetative genome replication (reviewed by Doorbar et al., 2012). However maintenance replication occurs in the typical S phase of basal keratinocytes and an effect of CTCF on the time of replication needs further investigation to draw conclusions on a possible role in episome replication during the maintenance phase.

Episome loss upon the abrogation of CTCF binding was also seen in EBV (Tempera *et al.*, 2010). They found that the latency specific promoter Qp was no longer protected from *de novo* methylation in a CTCF binding deficient mutant (Tempera *et al.*, 2010). This led to a slow and gradual DNA methylation that took 16 weeks until the promoter was completely silenced (Tempera *et al.*, 2010). In the samples tested in this study a significant loss of episomes was observed in samples of passage 14/15 or older. This corresponds to about 15 weeks of culturing and therefore correlates to the episome loss observed in EBV. However in HPV18 there are only two main promoters (early and late promoter) and one secondary promoter which is solely dedicated to the transcription of E8^E2C (Wang *et al.*, 2011). The early promoter could be silenced by methylation which would result in a general reduction in the expression of early genes and an inability of maintaining the viral genome. Also the promoter that is only used for the transcription of E8^E2C could be silenced by methylation since any change in transcription from this promoter could potentially cause genome loss.

In conclusion the known reasons of episome loss are changes in the expression levels of the proteins E2, E8^E2C, E6 and E7. There is a multitude of different mechanisms through which the abrogation of CTCF binding could have caused such changes. One of these mechanisms is the transcription regulation by interaction with the promoter either through targeting of enhancers and transcription factors or through recruiting of the RNA polymerase II to the promoter. CTCF could also alter the expression of these genes by altering methylation of the DNA, including the methylation of various E2 binding sites. This is interconnected with the third mechanism; the alteration of the splicing of viral transcripts. Alternative splicing could be achieved by CTCF-controlled methylation of splice sites or the stalling of the RNA polymerase II during transcription.

Considering the data presented in this study I suggest the interference with methylation to be the most likely solution to the episome loss. The slow accumulation of CpG methylation in response to abrogation of CTCF binding has already shown to be the cause of episome loss in EBV (Tempera *et al.*, 2010). Also the timing of the episome loss observed in EBV was very similar to the one seen in the experiments presented in this study. In EBV, a promoter was methylated whereas the conserved CTCF binding site in high risk HPV is located 3000 nucleotides away from the promoter for the early genes. It is uncertain if the protective function of CTCF can extend to the LCR but also methylation of the ORFs of the early genes could lead to epigenetic silencing of the early genes.

4.6 CTCF could regulate splicing in HPV

CTCF can influence splicing through two different mechanisms: the stalling of RNA polymerase II and the regulation of DNA methylation (Maunakea *et al.*, 2013, Shukla *et al.*, 2011). The stalling of the RNA polymerase II takes place when the polymerase encounters a CTCF-cohesin complex bound to the template DNA. This favours the assembly of a splicing complex upstream of the CTCF binding site, leading to the inclusion of weak upstream exons (Shukla *et al.*, 2011). The inclusion of weak upstream exons has been confirmed for the first 1000 nucleotides upstream of the CTCF binding site but it may extend further than that (Shukla *et al.*, 2011). The interplay of CTCF and methylation is rather complex since CTCF can protect DNA from *de novo* methylation and CTCF binding itself is abrogated by methylation (Guastafierro *et al.*, 2008, Zampieri *et al.*, 2012). The protection from *de novo* methylation has been shown to depend on a post-translational modification of CTCF called poly(ADP)ribosylation or PARlation (Caiafa *et al.*, 2009). PARlated CTCF can to protect an area of at least 2kb in the vicinity of its binding site from *de novo* methylation (Zampieri *et al.*, 2012, Davalos-Salas *et al.*, 2011, Fedoriw *et al.*, 2004, Szabo *et al.*, 2004).

The CTCF binding maps of HPV can be combined with transcription maps to draw conclusions as to how CTCF could influence splicing of particular transcripts. Figure 64 to Figure 68 show the transcript maps for HPV16R, HPV18 and HPV31 with all EMSA confirmed binding sites indicated as red vertical bars (reviewed by Van Doorslaer et al., 2013). Not all HPV transcripts could be determined so these binding maps do not show the complete transcriptome of HPV. The transcript of the essential protein E1 is missing in some of the maps, most likely to the low abundance of this transcript.

The strong and conserved CTCF binding site around nucleotide 3000 is the only binding site that could be confirmed in ChIP experiments using either HPV18 maintaining HFK cells or HPV16 maintaining W12 cells. Exon inclusion upstream of the CTCF binding site caused by RNA polymerase II stalling may play a role in HPV16 because the transcripts I, J and K contain a complete exon just upstream of the CTCF binding site. These are the only transcripts that can encode a truncated form of the E1 helicase; E1C. Thus E1C expression may depend on CTCF binding. Expression of E1C in HPV infected cells has not been confirmed and there is no information available on E1C so the transcripts encoding E1C may just offer improved or impaired translation of the other proteins encoded on these transcripts. Thus the expression of E2, E5, E6 and E7 may be altered by increased transcription of the transcripts I, J and K.

All high risk HPV genomes tested for CTCF binding have parts of the exon for the E1 transcript within 1000 bp upstream of the CTCF binding site. Thus abrogation of CTCF binding could have decreased the inclusion of the exon for E1, leading to reduced E1 expression. However only increased expression of E1 has been suggested to interfere with episome copy number by interference with tethering (Voitenleitner and Botchan, 2002). The dependency on E1 for maintenance replication may also be connected to the mode of replication in particular cell lines so there is the possibility that the reduction in E1 may have reduced episome copy if the episome replicated in a particular mode of replication, for example random choice replication.

Furthermore, post translational modifications of CTCF upon cellular differentiation could trigger the inclusion of the E1 exon later in the viral life cycle to support vegetative replication. Unfortunately, the expression of E1 protein in methylcellulose differentiated cells could not be analysed due to lack of functional antibody.

E1 expression in BPV1 could also be regulated by the strong CTCF binding site around nucleotide 3000 (Figure 68). However the widespread distribution of the other weak CTCF binding sites throughout the BPV1 genome could alter splicing in a multitude of ways which makes predicting particular mechanisms difficult.

Assuming that RNA polymerase II stalling has an effect on the splicing of mRNA sequences further upstream than 1000 nucleotides of the CTCF binding site, the presence of CTCF around nucleotide 3000 may alter the splicing of all viral transcripts. There is also the possibility that CTCF acts as a roadblock during transcription and delays the transcription of every single viral transcript through the stalling of RNA polymerase II. This delay in

transcription could be a regulatory mechanism that reduces viral gene expression in primary keratinocytes in order to reduce proliferation and immunogenicity while maintaining a relatively high copy number of viral episomes. However, the low risk specific strong and conserved binding site at 5400 would only downregulate the late transcripts since it is situated downstream of the early polyadenylation signal (Figure 67). Considering that the low risk E2, E6 and E7 are less potent in their functions compared to their high risk equivalents, a possible downregulation of these early proteins by CTCF may not be essential for low risk HPV which could make a CTCF binding site in this region obsolete (Hou *et al.*, 2002, Kovelman *et al.*, 1996, reviewed by Klingelhutz and Roman, 2012).

Also there is an exon 1100 nucleotides upstream of the CTCF binding site in low risk HPV so there is a chance that CTCF may promote the inclusion of this exon, resulting in the only mRNA transcript that is able to encode for L2. Thus the conserved CTCF binding site of low risk HPV may be involved in the splicing of late gene transcripts. However, suggesting a function for the CTCF binding sites that were only confirmed by EMSA is difficult since the putative effects regarding splicing and methylation if all these binding sites were occupied by CTCF could affect nearly the entire genome. As of now it is not known which of the EMSA confirmed binding sites are occupied by CTCF *in vivo* during the viral life cycle.

A potential function in protection from *de novo* methylation within a wide area around the strong and conserved CTCF binding sites would cover many HPV genes. The genes in an area of 2 kbp around the low risk specific binding site at nucleotide 5400 include L1 and L2. The strong and conserved binding site of high risk HPV around nucleotide 3000 could possibly protect E1, E2, E4 and E5 from *de novo* methylation. However these are just the proteins in the confirmed area of protection from *de-novo* methylation so the extension of protection to a wider area is possible (Szabo *et al.*, 2004, Fedoriw *et al.*, 2004).

Interestingly, western blotting of methylcellulose differentiated cells showed differences in the expression of E2 and E1^E4 (Figure 53). The lack of CTCF could lead to a change in methylation of these genes in the C3 mutant which could have resulted in decreased transcription through epigenetic silencing of these genes. Only the amount of E1^E4 was reduced in the C3 mutant compared to the wild type in methylcellulose differentiated cells. This effect varied between donors and was only significant in Georgie cells (Figure 53). However the data from raft cultures showed the same percentage of E1^E4 expressing cells in WT and C3 mutant (Figure 60). The reason for this may be a possible reduction of the

amount of E1^E4 per cell rather than a reduction in the number of E1^E4 expressing cells. Thus a possible reduction of E1^E4 expression in rafts may have gone unnoticed.

Nevertheless the increased and earlier expression of E2 in the C3 mutant upon differentiation contradicts the hypothesis of decreased gene expression due to methylation of the E2 gene. The intermediate passage number (passage 11) of the cells in which increased expression of E2 was observed leaves the option that accumulation of methylation may still downregulate E2 later in time. Instead, the lack of CTCF mediated enhancer blocking, looping or the lack of roadblock function could have increase E2 expression without accumulation of methylation.

Another interesting factor is the time frame in which *de novo* methylation takes place. In EBV it was shown that *de novo* methylation due to the lack of CTCF binding caused changes in gene expression only after 8 weeks of culturing, and complete silencing through increased methylation of particular sequences took 16 weeks (Tempera *et al.*, 2010). Methylcellulose differentiated cells of passage 11 were used for western blotting. This corresponds to 11 weeks of culturing.

The raft cells were generated from cells of passage numbers 6 and 10. This corresponds to 5 and 10 weeks of culturing, respectively. Another two weeks of culturing were needed to generate the rafts. Some changes in methylation could have occurred in this time frame but it is possible that a stronger effect would have been seen with cells of higher passage number. However cells with mutant genomes of higher passage number started losing significant numbers of genomes which may have been a result of progressing *de novo* methylation. The comparison of raft cultures from wild type and C3 cells with considerably different viral copy numbers would have introduced a bias into the data.

The weaker CTCF binding sites in HPV congregate in the ORFs for the late genes. If they are used *in vivo* it is likely that these binding sites could regulate the splicing of late gene transcripts. A protective function towards preventing *de novo* methylation is unlikely since it was already shown that none of these binding sites was occupied by CTCF in undifferentiated HFK or W12 cells maintaining HPV18 or 16, respectively. Thus these genes would not be protected against *de novo* methylation in the basal cells of the epithelium in which the virus can reside for decades.

All binding sites within the late genes showed an either weak or medium band shift in EMSA. The observation made by Essien *et al.* that low occupancy CTCF binding sites are more likely to be tissue specific supports the hypothesis that these binding sites may be occupied later in the viral life cycle when the cell differentiates (Essien *et al.*, 2009). Recently two novel E2^E4 splice transcripts of HPV18 were discovered in raft culture and CIN I-II biopsy (Tan *et al.*, 2012). These transcripts are called E2^E4-S and E2^E4-L and were shown to be upregulated upon keratinocyte differentiation. Translation of the E2^E4-L protein has been confirmed by the presence of a 23kDa protein in raft culture and CIN biopsy (Tan *et al.*, 2012). As yet their function remains unknown but the splicing pattern of the transcripts was analysed. The splice site used to generate the E2^E4-S transcript is situated at 2853, so just upstream of the CTCF binding site around nucleotide 3000. The splice site for E2^E4-L is situated just downstream of the CTCF binding site at nucleotide 3165 as seen in Figure 63 (Tan *et al.*, 2012). Thus the presence or absence of CTCF may be involved in the regulation of these two transcripts.

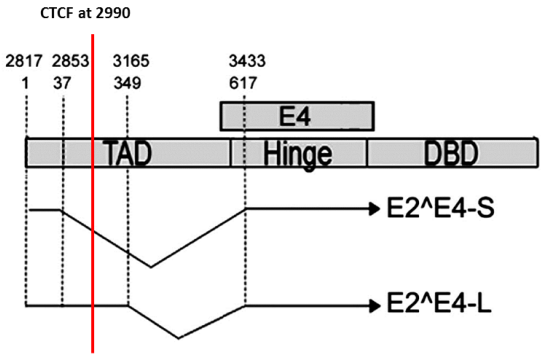


Figure 63) The E2 ORF of HPV18 with the transcripts of E4^E2-S and E4^E2-L
CTCF binding around nucleotide 3000 is indicated with a red line. It is located in-between the splice sites and may regulate the ratio of both splice variants. Figure of Tan *et al.*, modified with permission (Tan et al., 2012).

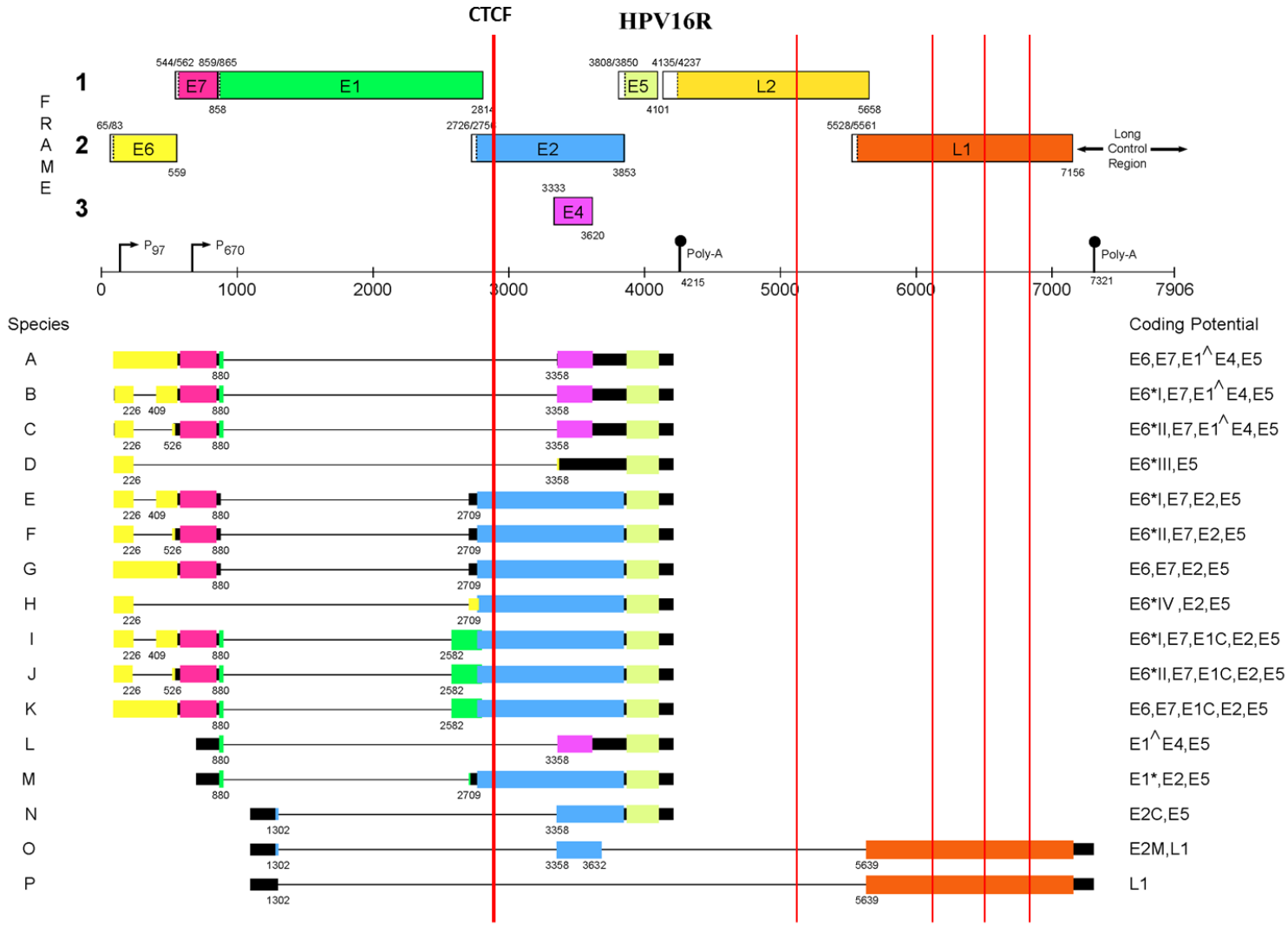


Figure 64) Transcription map of HPV16R with the EMSA confirmed CTCF binding sites

The strong and conserved CTCF binding site of high risk HPV around nucleotide 3000 (thick red bar) has also been confirmed in ChIP using undifferentiated W12 cells (unpublished data from Dr. Ian Groves from the Coleman laboratory, University of Cambridge, UK). The only exon within 1000 bp upstream of the conserved CTCF binding site around nucleotide 3000(thick red bar) encodes for E1C. All other EMSA confirmed binding sites are indicated with thin red vertical bars. Figure of Van Doorslaer *et al.*, modified with permission (Van Doorslaer *et al.*, 2013).

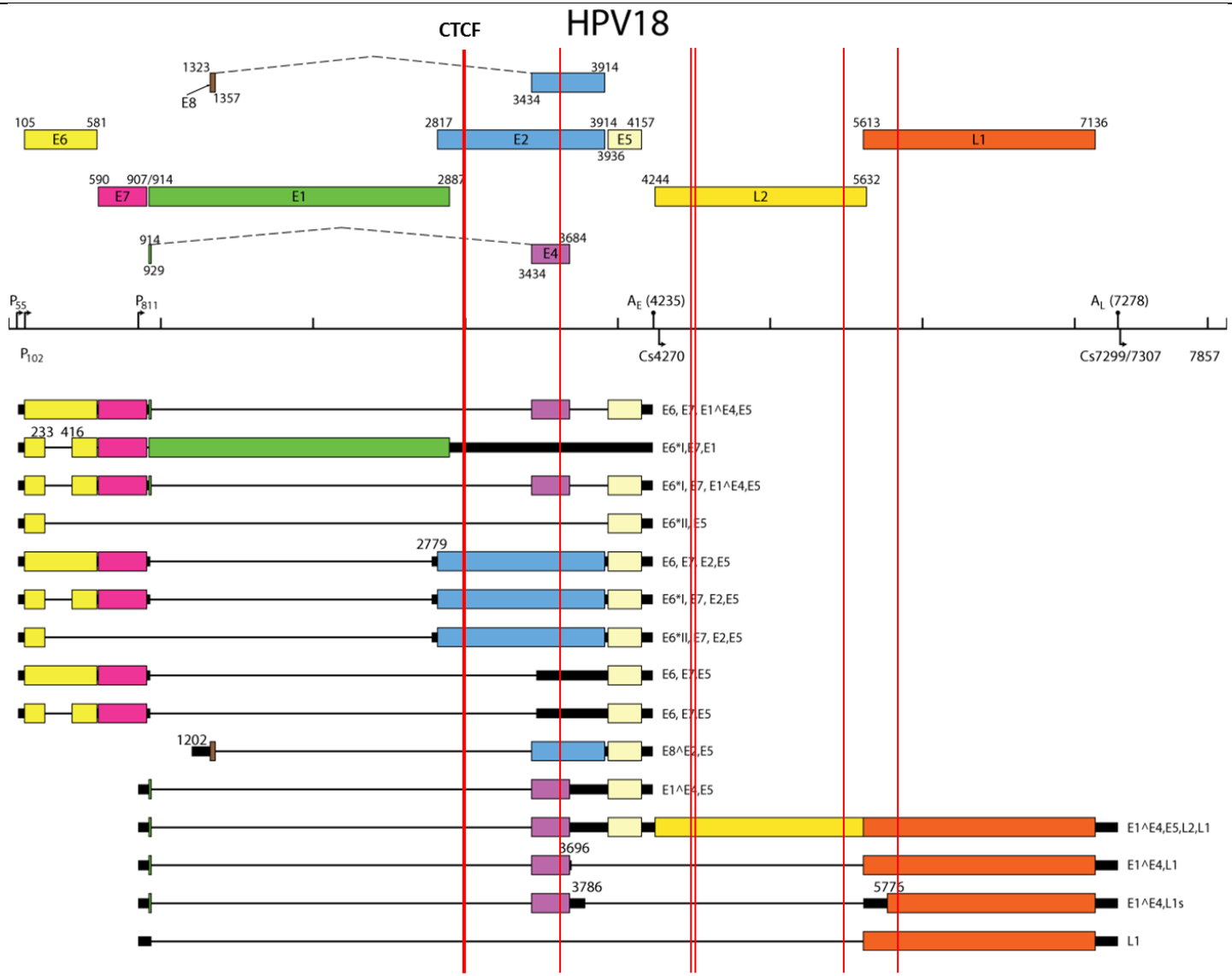


Figure 65) Transcription map of HPV18 with the EMSA confirmed CTCF binding sites

The strong and conserved CTCF binding site of high risk HPV at around nucleotide 3000 (thick red bar) was also confirmed in ChIP experiments with undifferentiated HFK cells. Only parts of the exon of E1 are within 1000 bp upstream of the conserved CTCF binding site around nucleotide 3000. All other EMSA confirmed binding sites are indicated with thin red vertical bars. Figure of Van Doorslaer *et al.*, modified with permission (Van Doorslaer *et al.*, 2013).

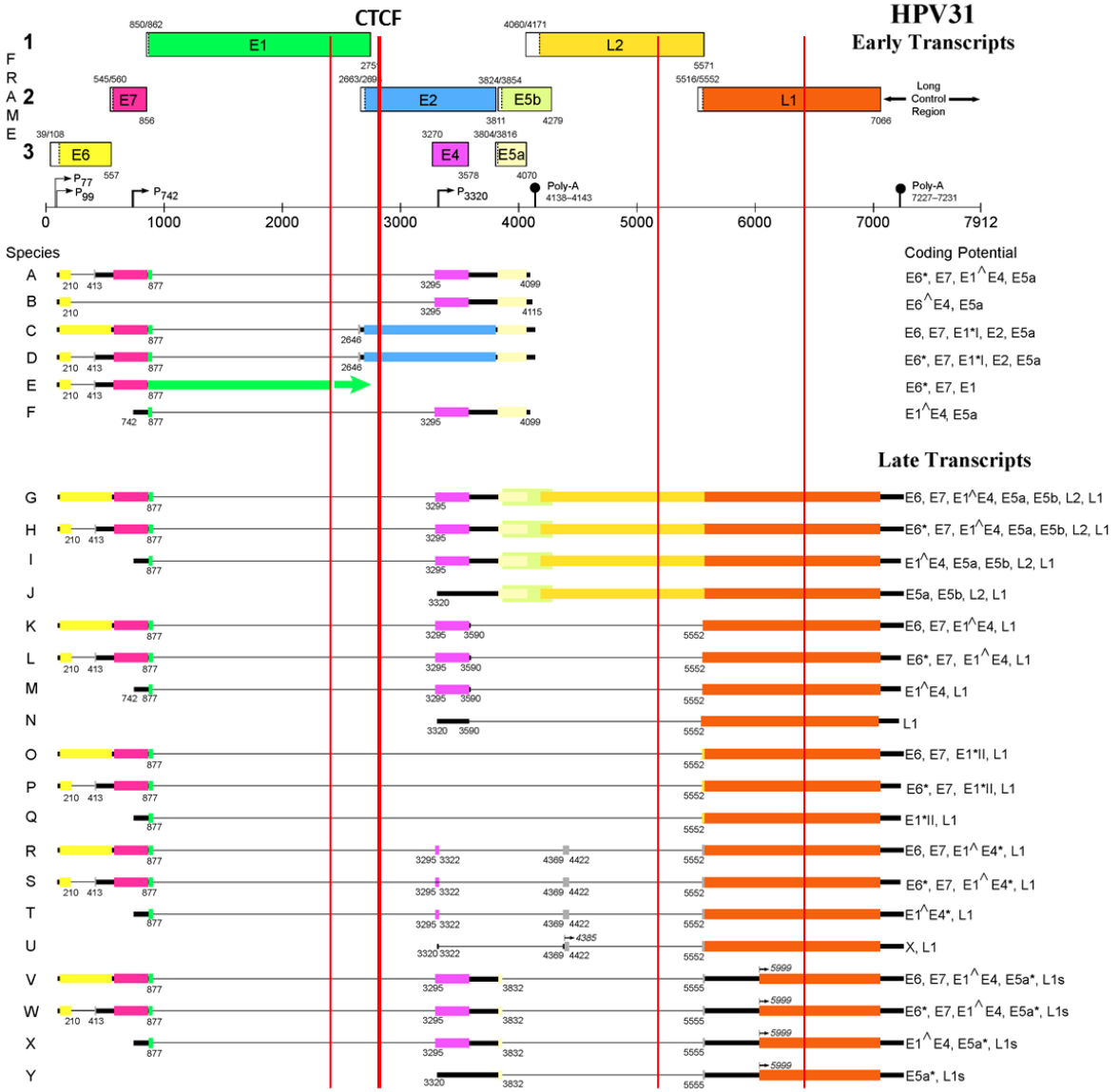


Figure 66) Transcription map of HPV31 with the EMSA confirmed CTCF binding sites

Only parts of the exon of E1 are within 1000 bp upstream of the conserved CTCF binding site around nucleotide 3000 (thick red bar). All other EMSA confirmed binding sites are indicated with thin red vertical bars. Figure of Van Doorslaer *et al.*, modified with permission (Van Doorslaer *et al.*, 2013).

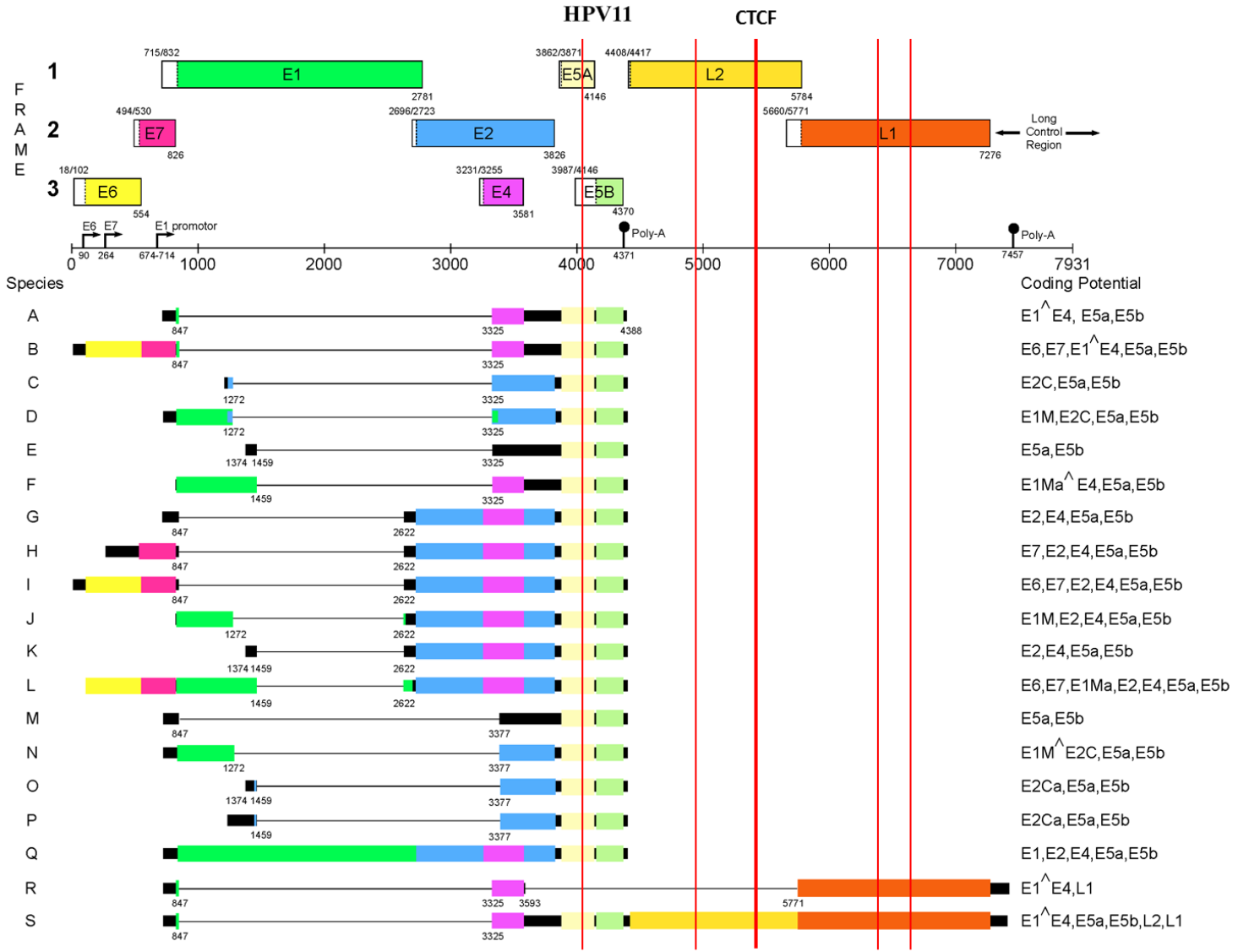


Figure 67) Transcription map of HPV11 with the EMSA confirmed CTCF binding sites

The only exon within 1000 bp upstream of the conserved CTCF binding site around nucleotide 5400 (thick red bar) is the exon encoding L2. All other EMSA confirmed binding sites are indicated with thin red vertical bars. Figure of Van Doorslaer *et al.*, modified with permission (Van Doorslaer *et al.*, 2013).

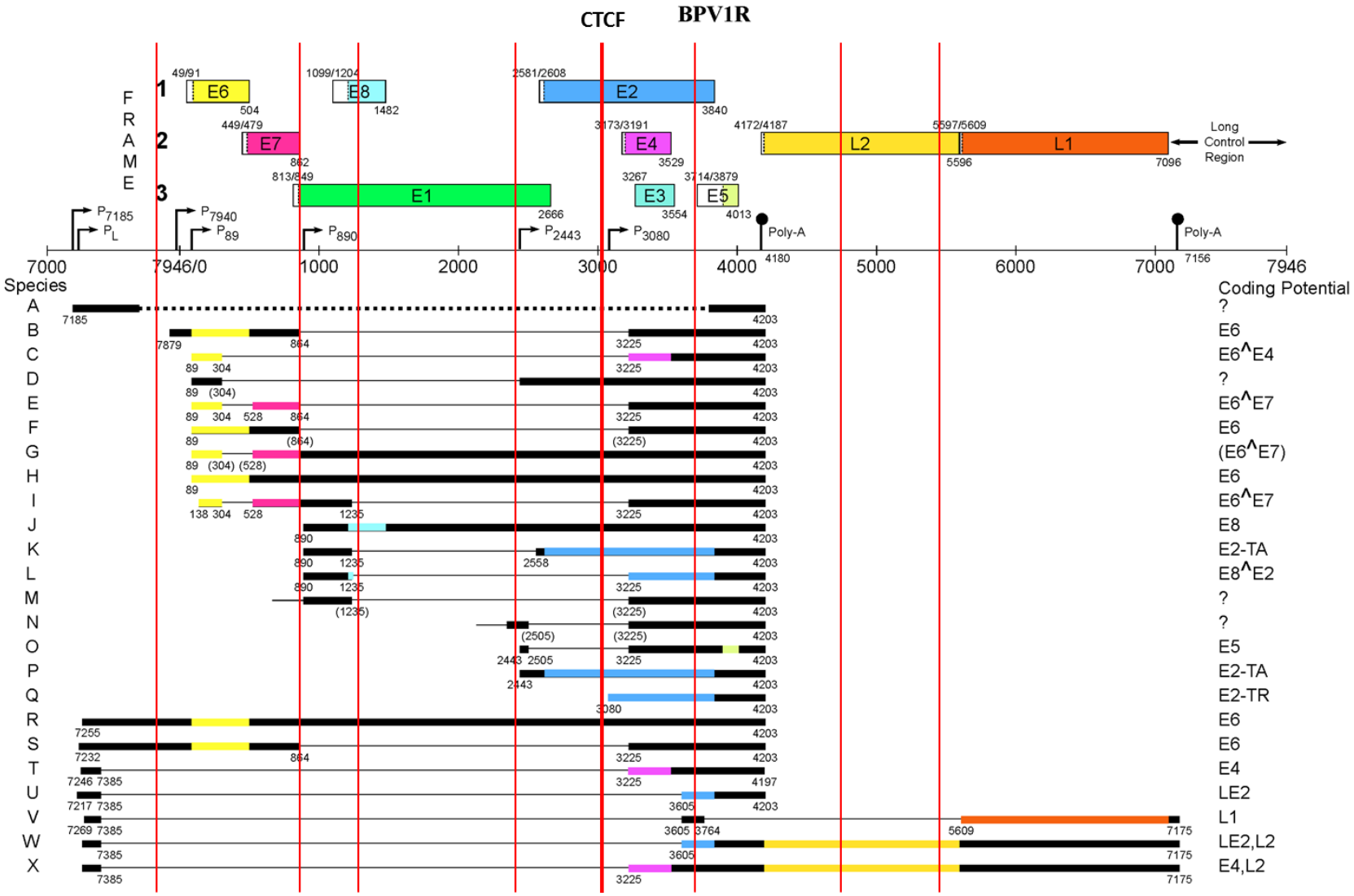


Figure 68) Transcription map of BPV1R with the EMSA confirmed CTCF binding sites

Only parts of the exon of E1 are within 1000 bp upstream of the CTCF binding site around nucleotide 300 (thick red bar). All other EMSA confirmed binding sites are indicated with thin red vertical bars. Figure of Van Doorslaer *et al.*, modified with permission (Van Doorslaer *et al.*, 2013).

4.7 The role of CTCF in enhancer blocking, targeting of transcription factors and recruiting of RNA polymerase II

CTCF is able to regulate gene expression though a multitude of different mechanisms including the blocking of enhancers, the targeting of transcription factors and recruiting of the RNA polymerase II to promoters (Hark *et al.*, 2000, Chernukhin *et al.*, 2007).

For the H19/Igf2 locus it was shown that CTCF needs to be located between a promoter and enhancer to block the enhancer from acting on the promoter (Banerjee *et al.*, 2001, Kurukuti *et al.*, 2006, Hark *et al.*, 2000). This function was shown to depend on PARlation of CTCF (Farrar *et al.*, 2010, Yu *et al.*, 2004). The blocking of an enhancer often results in re-targeting of the enhancer so that it acts on an alternative promoter (reviewed by Wallace and Felsenfeld, 2007, Ling *et al.*, 2006, Kanduri *et al.*, 2000). Similarly, the formation of loops can result in re-targeting of transcription factors (reviewed by Holwerda and de Laat, 2013). The major transcription factor in HPV gene regulation is E2 which is known to preferably bind within the LCR (reviewed by McBride, 2013). The entire genome of BPV1 has been thoroughly screened for E2 binding sites, resulting in the identification of 17 binding sites, most of which are situated within the LCR (Li *et al.*, 1989). In the last two decades the majority research in HPV gene regulation has focused on the LCR while the E2 binding sites outside the LCR have been neglected (reviewed by McBride, 2013). The results

Area	E2 binding site no.	Location	CTCF binding fragment within 100 bp
LCR	1	7203	no
	2	7365	no
	3	7408	no
	4	7459	no
	5	7510	yes
	6	7592	yes
	7	7621	yes
	8	7635	yes
	9	7761	yes
	10	7781	yes
	11	7896	yes
	12	16	no
Coding region	13	855	yes
	14	1125	yes
	15	2396	yes
	16	2921	yes
	17	3088	yes

Table 39) Locations of E2 binding sites in BPV1
Data from Li *et al.* reproduced with permission (Li et al., 1989)

presented in this study support the idea that E2 binding sites outside the LCR may still play important roles in PV gene regulation. As seen in Table 39 and Figure 69, all of the E2 binding sites in BPV1 situated outside the LCR co-localise with EMSA confirmed CTCF binding sites. Also there are many CTCF binding sites that are situated in-between the LCR and E2 binding sites and. Thus the E2 proteins bound outside the LCR could be blocked from acting on the LCR by looping mediated by CTCF. Using this mechanism, CTCF could target E2 action on particular promoters.

Also every promoter in BPV1 that is located outside of the LCR co-localises with CTCF and CTCF was shown to be able to recruit RNA polymerase II and induce transcription (Figure 69) (Chernukhin *et al.*, 2007).

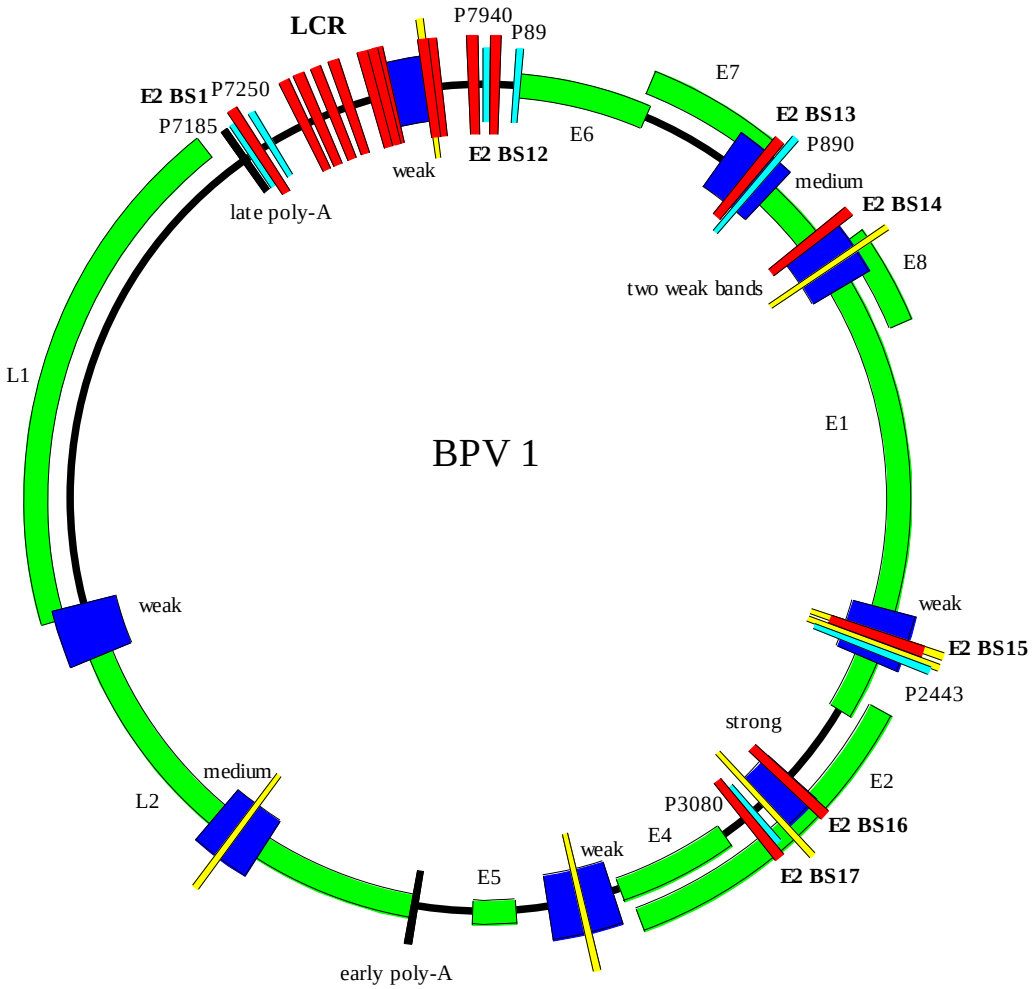


Figure 69) Co-localisation of CTCF with E2 and promoters in BPV1
CTCF binding fragments are shown in blue, promoters are teal and E2 binding sites red. All E2 binding sites and promoters outside the LCR co-localise with CTCF binding sites.

Yet another function of CTCF may play a role in the co-localisation with E2. CTCF is known to recruit PARP1 and induce PARlation of proteins in its vicinity (Guastafierro *et al.*, 2008). Via this mechanism CTCF inhibits *de novo* methylation by PARlating the DNA methyltransferase DNMT1, resulting in deactivation of this enzyme (Zampieri *et al.*, 2012). In contrast to this, HPV18 E2 was shown to only be transcriptionally active if it is PARlated (Lee *et al.*, 2002). Thus CTCF could ensure E2 PARlation to induce transcription. Since the protection from de-novo methylation through CTCF-induced PARlation of DNMT1 covers an area of at least 2 kbp around the CTCF binding site, also CTCF-induced PARlation of E2 could possibly extend over this area (Fedoriw *et al.*, 2004, Szabo *et al.*, 2004). Furthermore it is not known if there is an E2 binding around the CTCF binding site at nucleotide 3000 in high risk HPV as it is seen in BPV1, so there may be an E2-PARlation target in close vicinity to the strong and conserved CTCF binding site of high risk HPV. However the association of CTCF with E2 and promoters is only seen in BPV1 so this function may have a bigger impact on BPV1 than on HPV. The region where most of the promoters outside the LCR are located in BPV1 has been shown to be devoid of CTCF binding sites in HPV. Most HPV have either no or few promoters located outside the LCR and there is no CTCF binding site predicted in the LCR of these viruses. Nevertheless the CTCF binding site in BPV1 that is located within the LCR was not predicted either so there is the chance that CTCF could still be recruited to the LCR of HPV at some point during the viral live cycle.

Another transcription factor, the estrogen receptor, has been shown to be located close to nucleotide 3000 in HPV16 (Kitasato *et al.*, 1997). Its position has been defined to be between the nucleotides 3142 and 3255 (Kitasato *et al.*, 1997). Thus CTCF is located in between this transcription factor and the LCR. Closer investigation of an isolated fragment containing this estrogen receptor binding site showed that it had no effect gene regulation (Kitasato *et al.*, 1997). This isolated DNA fragment did not contain the CTCF binding site around nucleotide 3000 (Kitasato *et al.*, 1997). CTCF is known to facilitate the regulatory function of the estrogen receptor when it is bound in close proximity. This may have been the reason why no effect was seen (Kitasato *et al.*, 1997, Ross-Innes *et al.*, 2011). There is an estrogen response element in the LCR which was shown to be involved in carcinogenesis (Arbeit *et al.*, 1996, Gariglio *et al.*, 2009). The location of CTCF between the oestrogen receptor binding site and this estrogen response element supports the hypothesis that CTCF could possibly block the estrogen receptor from acting on the estrogen response element. Thus a role for CTCF in HPV induced carcinogenesis is possible. This is supported

by the fact that the HPV18 positive HeLa cells that originated from a high grade cervical carcinoma were shown not to bind CTCF around nucleotide 3000 of integrated genomes (reviewed by Johannsen and Lambert, 2013). Since oestrogen promotes HPV induced carcinogenesis it is possible that the abrogation of CTCF exposes the estrogen receptor to the estrogen response element resulting in altered viral genome expression that promotes cancer. It would be interesting to see how the cell lines containing wild type and mutant HPV18 respond to stimulation by oestrogen. However the presence of the estrogen receptor binding site is so far only confirmed for HPV16.

4.8 The viral life cycle is altered when CTCF binding is abrogated

The comparison of raft cultures and western blots from cells maintaining wild type HPV18 or C3 mutant revealed distinct differences between these two viruses. Rafts grown from mutant C3 maintaining cells were thicker compared to the rafts from wild type maintaining cells and they resembled pre-cancerous lesions. Immunohistochemical staining of the markers BrdU (S-phase) and Cyclin B1 (G2 entry) revealed that cells in the upper layer of the C3 raft were dividing more often compared to the wild type maintaining cells. This increase in proliferation was only seen in differentiating cells whereas the growth curves of undifferentiated cells maintaining wild type and mutant HPV18 looked identical (unpublished data from Dr. Jo Parish). Thus viral gene regulation in advanced stages of the viral life cycle was affected by the mutations that abrogated CTCF binding around nucleotide 3000.

Analysis of E1^E4 using immunohistochemistry revealed the same pattern in wild type and mutant maintaining cells and therefore contradicted the result from methylcellulose differentiated cells which showed decreased E1^E4 expression in the C3 mutant. A reason for this may be a possible reduction of the amount of E1^E4 per cell rather than a reduction in the number of E1^E4 expressing cells. Thus a probable reduction of E1^E4 expression in rafts may have gone unnoticed. Another explanation is that the generally high level of E1^E4 expression per cell may have led to the inability to detect subtle changes in E1^E4 expression due to limiting factors like the availability of antibody. On the other hand methylcellulose differentiated cells do not fully resemble the typical differentiation stages. Hence the different results of both differentiation methods could perhaps be artefacts. Also the possibility of episome loss in the C3 mutant could play a role. Since E1^E4 expression is

generally very strong, frequent transcription may occur and the reduction in template DNA may have decreased the overall expression of E1^E4. Cells of passage 11 were used for differentiating cells in methylcellulose. Those cells showed a reduction in C3 episome number of approximately 60 % which could explain the reduction in E1^E4. On the other hand cells of low passage number (passage 6), which showed no reduction in episomes, have been used for generating the rafts. The rafts did not show a difference in E1^E4 expression between WT and mutant.

However, decreased expression of E1^E4 protein provides some limited explanation for the phenotype of C3 mutant raft cultures because E1^E4 is one of the proteins involved in creating the pseudo S-phase by stopping cell cycle progression while E6 and E7 promote cell cycle progression (Knight et al., 2004, Knight et al., 2011, Knight et al., 2006, Davy et al., 2002, Davy et al., 2006). A reduction in E1^E4 is likely to allow the viral oncoproteins to push the cells into mitosis which would explain the increased number of dividing cells in the C3 mutant compared to the wild type (reviewed by McBride, 2008, Davy et al., 2002).

Staining for loricrin revealed another distinct difference between wild type and mutant rafts (Figure 61). In the wild type rafts, cells in the keratinous layer flattened out and formed a typical solid and keratinous layer that protects the skin. Cells maintaining the C3 mutant did not flatten out and rather than forming a protective layer of keratin, round keratinous cells were seen that left gaps between them. Morphological differentiation in the C3 mutant appeared to be impaired whereas expression of the differentiation marker loricrin remained intact. Morphological abnormalities in HPV infected cells are often the result of E5 and E6 expression (Crusius *et al.*, 2000, Crusius *et al.*, 1998, Krawczyk *et al.*, 2008, Stoppler *et al.*, 1996, Straight *et al.*, 1993). Since the expression of E6 was shown to be constant throughout the differentiation in methylcellulose, the factor responsible for abnormal morphology in the C3 mutant cells is likely to be E5. Thus the expression of E5 may be upregulated in the C3 mutant compared to the wild type.

The growth pattern seen in the mutant rafts resembled pre-cancerous lesions that usually occur when the negative control over the early genes in HPV is lost (Lowy and Schiller, 2006). However the expression of the viral oncogenes E6 and E7 was shown to be normal in methylcellulose differentiated cells (Figure 54).

Instead the expression of E2 occurred earlier. At low expression levels E2 stimulates oncogene expression but at higher levels it is able to inhibit expression of the same oncogenes (Steger and Corbach, 1997). However the levels of E6 and E7 were normal in

methylcellulose differentiated cells (Figure 54). Thus the increase in proliferation seen in the C3 mutant is likely to be caused by a factor other than the overexpression of E6 and E7. The overexpression of E5 is a possible explanation since this protein can also stimulate proliferation and enhances the efficiency of immortalisation by E6 and E7 (Straight *et al.*, 1993, reviewed by Klingelhutz and Roman, 2012, Maufort *et al.*, 2010, Stoppler *et al.*, 1996).

In conclusion the abrogation of CTCF around nucleotide 3000 disrupts events in the mid- and late stages of the viral live cycle and promotes hyperplastic morphology of raft cultures despite typical expression of viral oncogenes E6 and E7 in methylcellulose differentiated cells. Thus CTCF has the potential to be an important regulatory factor in the HPV life cycle and cancer progression.

4.9 A potential role of CTCF in genome integration, episome positioning and DNA looping

CTCF has been shown to be involved in the general arrangement of DNA through mechanisms like the formation of loops, bridges and potentially the positioning of DNA inside the nucleus (reviewed by Zlatanova and Caiafa, 2009a). Since the nucleus is subdivided into transcriptionally active and inactive compartments, CTCF could potentially regulate viral transcription by positioning the episome in a location that favours or inhibits transcription (reviewed by Zlatanova and Caiafa, 2009a). It has been shown that small and gene rich chromosomes tend to be situated more centrally in the nucleus (Bolzer *et al.*, 2005, reviewed by Cremer and Cremer, 2001). Hence confirmation of a CTCF binding site in the HPV genome opens up the possibility that CTCF may facilitate positioning of the HPV genome to transcriptionally active regions of the nucleus. However more research is needed to investigate this hypothesis.

DNA looping has already been confirmed in papillomaviruses; however this looping is presently attributed to E2 function. The E2 protein was shown to form loops in the viral episome by dimerisation through its N-terminal domain in HPV16, HPV11 and BPV1 (Hernandez-Ramon *et al.*, 2008, Sim *et al.*, 2008, Knight *et al.*, 1991, Antson *et al.*, 2000). Shorter forms of E2 were shown to be unable to form DNA loops suggesting that competition with shorter forms of E2 for E2 binding sites regulates loop formation (Knight *et al.*, 1991).

Interestingly, occupation of some E2 binding sites seems to depend on occupation of another E2 binding site which may be a result of looping. It was shown in BPV1 that mutation of E2 binding site 17 also resulted in abrogation of E2 binding at binding site 16 (Knight *et al.*, 1991). Both of these E2 binding sites are located outside the LCR. Unfortunately, the E2 binding sites outside the LCR in other HPV have not been published so it is hard to say if CTCF also co-localises with these E2 binding sites. It is worth noting that CTCF is not known among the direct interaction partners of E2, but there is at least one common interaction partner of CTCF and E2 which is the Poly [ADP-ribose] polymerase 1 (PARP1) (reviewed by McBride, 2013, Zampieri *et al.*, 2012). Furthermore the looping function of E2 in HPV16 was seen using affinity purified E2 protein, thus a contribution of CTCF to E2 dependent looping seems unlikely (Hernandez-Ramon *et al.*, 2008).

Nevertheless, the presence of CTCF binding sites in the HPV genome introduces the possibility of a second factor in HPV DNA looping. Potential loops from the high risk specific binding site at nucleotide 3000 to the LCR are especially interesting because of the presence of an estrogen receptor binding site in the vicinity of the CTCF binding site and an estrogen response element in the LCR (Kitasato *et al.*, 1997). Thus CTCF could form a loop to the LCR to activate transcription via the estrogen response element. This is supported by the fact that the transcription inducing function of the estrogen receptor can depend on a CTCF binding site in its vicinity (Ross-Innes *et al.*, 2011). Furthermore the presence of oestrogen and corticosteroids has been shown to promote HPV induced carcinogenesis (Arbeit *et al.*, 1996, Gariglio *et al.*, 2009). Thus it can be hypothesised that CTCF can possibly regulate viral gene transcription via the estrogen receptor through looping and could therefore be an important factor in carcinogenesis. However, while the oestrogen responsiveness of HPV-induced carcinogenesis is confirmed, the presence of a binding site for the estrogen receptor outside the LCR has only been confirmed in HPV16 so far (Arbeit *et al.*, 1996, Kitasato *et al.*, 1997).

The integration of the viral genome into the host genome is an important factor for HPV-induced carcinogenesis. Nearly all of HPV18 induced tumours of the cervix contain integrated viral DNA (Woodman *et al.*, 2003). Also in HPV16 induced tumours the prevalence of integrated genomes is high since only 30 % of these tumours contain exclusively episomal DNA (Badaracco *et al.*, 2002). The HPV genome was found to integrate semi-randomly into the host DNA, most often at common fragile sites (Ziegert *et al.*, 2003, Thorland *et al.*, 2003).

CTCF is known to form physical connection between distant DNA loci in trans (Zhao *et al.*, 2006, reviewed by Zlatanova and Caiafa, 2009a). These connections are called bridges and bring the otherwise distant loci in close proximity. Bringing viral and host DNA in close proximity could possibly contribute to the integration of the viral genome into host DNA. Also the looping function of E2 may play a role in this mechanism; however the function of E2 looping has so far only been shown in cis (Knight *et al.*, 1991).

The integration sites on the viral genome found in cervical tumours follow a loose pattern as seen in Figure 70. The most common carcinogenic integrants contain an intact LCR as well as the complete E6 and E7 ORFs (Cone *et al.*, 1992, Schwarz *et al.*, 1985). Also the open reading frame for E2 is commonly interrupted or inactivated (Pett and Coleman, 2007).

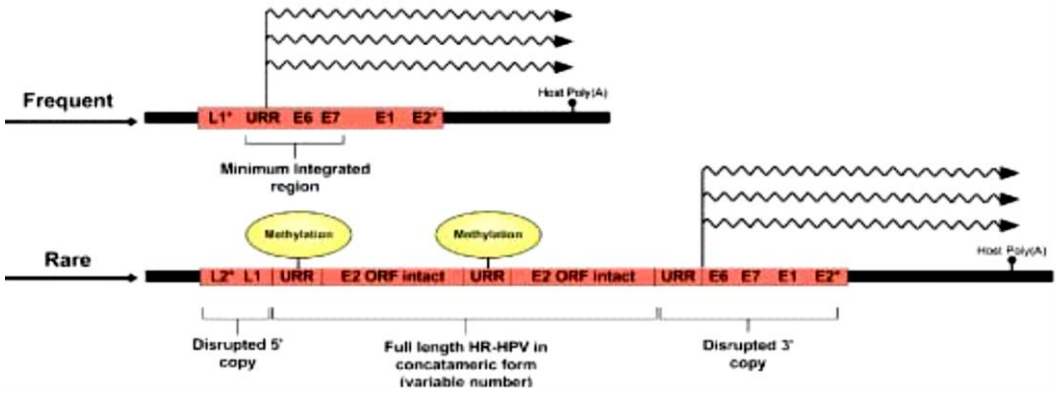


Figure 70) The integration pattern of HPV16
The LCR as well as the ORFs for the viral oncoproteins are often intact in integrants while the E2 ORF is disturbed. Thus the negative regulator of the viral oncogenes is no longer transcribed as episomes are lost. On rare occasions fragments of multiple episomes were found to be integrated into the host genome and these often have intact E2 ORFs. However methylation of the LCR regulating these ORF can abrogate E2 transcription so that viral oncogenes can be overexpressed regardless. Figure of Pett *et al.*, reproduced with permission (Pett and Coleman, 2007).

However the most common integrations overall, independently of their potential to cause cancer, occur in the E1 and L1 ORFs (Dall *et al.*, 2008, Wang *et al.*, 2013). The data on this is partly contradictory since another study claims that most integrations were observed within the L2 region (Li *et al.*, 2013). The integration sites determined by Dall *et al.* can be seen in Figure 71. Interestingly the integration sites discovered in this study are often situated within 1000 nucleotides of the strong and conserved CTCF binding site around nucleotide 3000. However more research is needed to determine if bridge formation by CTCF occurs in HPV and if bridge formation plays a role in integration of the viral genome.

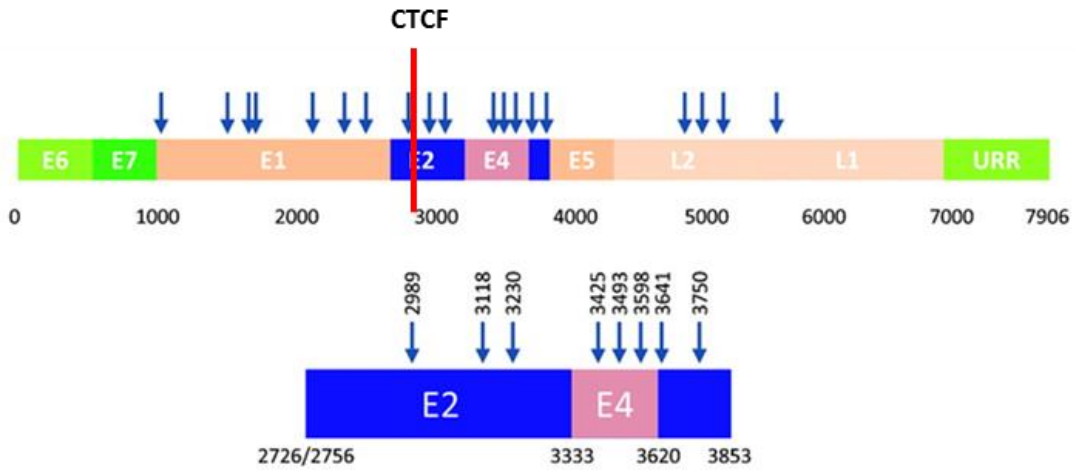


Figure 71) Integration sites of the HPV16 genome determined by Dall *et al*
The red bar marks the strong and conserved CTCF binding site. Most integrations (blue arrows) occurred within 1000 nucleotides of this binding site. Figure of Dall *et al.*, modified with permission (Dall et al., 2008).

Also the CTCF mediated protection from *de novo* methylation can be a factor in the integration of viral genomes since it was shown that mutation of various E2 binding sites in HPV31 strongly promotes integration (Stubenrauch *et al.*, 1998). The methylation of E2 binding sites may have the same effect since methylation abrogates E2 binding (Chaiwongkot *et al.*, 2013, Kim *et al.*, 2003, Thain *et al.*, 1996). The loss of CTCF may open up the viral genome to *de novo* methylation and could contribute to the integration process this way. It is interesting to note, that the HPV18 transformed HeLa cells no longer bind CTCF around nucleotide 3000 of their integrated HPV18 DNA (reviewed by Johannsen and Lambert, 2013). This is supported by the observation that integrated viral DNA is often methylated whereas actively replicating and transcribing viral DNA is generally unmethylated (Turan *et al.*, 2006, Chaiwongkot *et al.*, 2013).

4.10 Clinical implications

The data presented in this study gives rise to further research that has the potential to change diagnostic, prognosis and treatment of HPV infections.

The abrogation of CTCF binding around nucleotide 3000 in HPV was shown to increase cell proliferation and is suggested to open up E2 binding sites for methylation. Methylation of E2 binding sites abrogates E2 binding and often results in overexpression of the viral oncogenes, thus promoting cancer (Cheung *et al.*, 2013). The abrogation of E2 binding has

also been shown to promote integration of the viral genome which is an important factor that contributes to carcinogenesis (Stubenrauch *et al.*, 1998). It is interesting to note that HPV18 transformed HeLa cells do not bind CTCF at nucleotide 3000 in their integrated copies of HPV18 (reviewed by Johannsen and Lambert, 2013). All these observations indicate that the binding of CTCF could have an important impact on HPV-induced carcinogenesis. Thus CTCF binding or the methylation of the CTCF binding sites has the potential to give information on the likelihood of progression to cancer. If the abrogation of CTCF binding results in a high chance of a cervical neoplasm to progress to cancer this information could help clinicians to decide a suitable treatment strategy. However further research is needed to more closely investigate the role of CTCF in methylation protection, bridging and general disease progression before a diagnostic potential for CTCF in HPV-induced disease can be confirmed. Also it has to be made sure that effect seen upon abrogation of CTCF binding is not due the new potential binding site for GATA3 that was formed through the mutations introduced.

HPV episome stability has been suggested to be a possible drug target and there are studies that aim to clear an HPV infection by interfering with episome maintenance (Edwards *et al.*, 2011). Since CTCF has been shown to be involved in episome maintenance it could be another factor that needs to be considered for these kinds of treatment. However using CTCF as a direct drug target may be problematic due to the major impact on host cell gene expression. Nevertheless, targeting an interacting protein of CTCF or the CTCF binding site itself using gene-targeting medicine opens up the possibility of treating an HPV infection by impairing episome maintenance or viral gene expression through interference with CTCF function. Possible risks like an increase in cell proliferation have to be considered but could possibly be outweighed by increased immunogenicity of the host cell due to altered viral gene expression.

4.11 Comparison of HPV with other viruses that utilise CTCF

The role of CTCF has been studied in EBV, HVS, HSV and KSHV (Kang *et al.*, 2013, Tempera *et al.*, 2010, Ertel *et al.*, 2012, Zielke *et al.*, 2012). These viruses have a number of similarities to HPV but also distinct differences. They are all episomally maintained DNA viruses with a circular genome and distinct stages in their life cycles. In the latent stage they

are maintained at a stable copy number, undergo licensed replication and attach to host chromosomes via a tethering mechanism.

The maintenance of some of these viruses has been shown to depend on CTCF (Zielke *et al.*, 2012, Tempera *et al.*, 2010). However the reason for episome loss in response to abrogation of CTCF binding was suggested to be the downregulation of latency associated transcripts instead of a direct interaction of CTCF with the tethering mechanism (Zielke *et al.*, 2012, Tempera *et al.*, 2010). In the case of EBV the downregulation of latency transcripts was a gradual process that took 16 weeks until the latency transcript were not detectable any more (Tempera *et al.*, 2010). HPV genomes in the C3 mutants were lost within a similar time span making it possible, like in EBV, that slowly progressing *de novo* methylation of HPV DNA caused downregulation of the early genes and eventually the loss of viral episomes.

Also a role for CTCF in maintaining virus latency was discovered in all other viruses (Kang *et al.*, 2013, Tempera *et al.*, 2010, Ertel *et al.*, 2012, Zielke *et al.*, 2012). There is some evidence that a latency phase of HPV may exist (reviewed by Doorbar, 2013). Most of this evidence is based on animal models but there are some observations of recurrent HPV-related disease in immunosuppressed humans (Ozsaran *et al.*, 1999, Paternoster *et al.*, 2008). According to animal models, this putative latency stage is suggested to involve a reduction in the copy number of viral episomes per cell (Maglennon *et al.*, 2011, Nicholls *et al.*, 1999, reviewed by Doorbar, 2013). The abrogation of CTCF binding in HPV18 caused a reduction in viral episomes and it is likely that each cell retained a minimal number of episomes since the viral oncoproteins are needed for the immortalisation and transformation of the primary keratinocytes used to generate the cell lines. Thus there is a chance that the abrogation of CTCF binding around nucleotide 3000 could possibly induce a slow transition of the HPV activity into the latent stage.

One of the major differences between HPV and the other viruses is the size of the episome. The genomes of EBV, HVS, HSV and KSHV are about 20 times the size of the HPV genome. This indicates that long range interactions mediated by CTCF as well as the formation of chromatin borders may be more common to these viruses compared to HPV. Epigenetic roles of CTCF in these viruses have already been confirmed in terms of the maintenance of transcriptionally active or inactive domains as well as regulation of the methylation pattern (Ertel *et al.*, 2012, Kubat *et al.*, 2004b, Kubat *et al.*, 2004a, Tempera *et al.*, 2010). Despite the small size of the HPV genome, the regulation of methylation could have an important

impact on the HPV life cycle. The region on the HPV genome that encodes L1 and L2 is transcriptionally silent until the late stages of the viral life cycle. Thus epigenetic silencing of the late genes by histone modifications could potentially be maintained by the CTCF. The expression of late genes has not been tested in the C3 mutant. The high risk specific CTCF binding site within the E2 ORF upstream of the late genes is positioned in a similar fashion to the CTCF binding site of HSV that regulates the epigenetic silencing of the late genes which is situated within the ORF of the latency associated transcripts (Ertel *et al.*, 2012).

Furthermore, differences in the locations of the CTCF binding sites between HPV and the other DNA viruses studied have been observed. In HPV, CTCF binding sites were neither predicted nor confirmed in the LCR whereas CTCF binding sites of the other viruses accumulate within the LCR. However in BPV1 there is definite association of CTCF binding sites with promoters outside the LCR, and one weak binding site was found by EMSA within the LCR of BPV1. This indicates that the function of CTCF in BPV1 could be related to direct initiation of transcription by recruiting RNA pol II or by facilitating the PARlation of E2. For HPV it is more likely that the function of CTCF may be long ranged or related to enhancer blocking and epigenetics rather than a direct initiation of transcription.

CTCF has been shown to block enhancers, form loops and regulate splicing in other viruses, which gives further insight into the potential roles of CTCF in the HPV life cycle (Amelio *et al.*, 2006, Ertel *et al.*, 2012).

Accordingly, the estrogen receptor around nucleotide 3000 in HPV16 may be re-targeted through CTCF mediated looping so that it cannot act on the estrogen response element in the LCR (Arbeit *et al.*, 1996). The CTCF binding site is situated between the estrogen-receptor and the estrogen response element in the LCR, making it possible that CTCF binding to this site does indeed block transcriptional activation. A role of CTCF binding in viral mRNA splicing has been confirmed in KSHV (Kang *et al.*, 2013). The abundance of differently spliced transcripts upstream of the CTCF binding site around nucleotide 3000 in HPV makes tight regulation of splicing necessary, which could be mediated by CTCF. On the other hand the CTCF binding site in this location could also downregulate the transcription of viral early genes through RNA polymerase II stalling to ensure viral persistence in basal cells.

4.12 Outlook on the future

This study is the first analysis of the role of CTCF in the papillomavirus life cycle. However many questions remain due to the multitude of different functions CTCF can provide.

So far, the binding pattern of CTCF in undifferentiated keratinocytes has been confirmed but it is uncertain how the binding pattern will change in differentiated cells. For HPV-induced cancer tissue the only data available on CTCF binding to integrated HPV genomes shows absence of CTCF binding around nucleotide 3000. However this data is from HeLa cells which are known to have many genetic abnormalities (reviewed by Johannsen and Lambert, 2013). The determination of the CTCF binding pattern throughout the cell cycle in HPV positive cancer tissues and in latently infected tissue will give a deeper insight into the role of CTCF HPV gene regulation and cancer development.

The function of CTCF regarding methylation of the HPV genome is another important question. Methylation is of vital importance in HPV gene regulation and studies of EBV show that CTCF has the potential to protect episomal viruses from *de novo* methylation (Tempera *et al.*, 2010). Methylation was also shown to contribute to integration of the viral genome and the upregulation of viral oncogenes which raises questions regarding the role of CTCF in cancer development (Turan *et al.*, 2006, Chaiwongkot *et al.*, 2013, Stubenrauch *et al.*, 1998, Thain *et al.*, 1996). Thus unravelling how CTCF influences viral methylation has the potential to improve the understanding of HPV gene regulation considerably.

The possible formation of DNA bridges between HPV genomes and host DNA is another topic to investigate since it may promote integration and could affect gene regulation. Loops in HPV DNA can also be formed by E2; however the regulatory nature of this mechanism is poorly understood and E2-looping has only been confirmed in cis (looping/bridging in trans may still occur) (Hernandez-Ramon *et al.*, 2008, Sim *et al.*, 2008, Knight *et al.*, 1991, Antson *et al.*, 2000). CTCF could be another player involved in the looping of HPV DNA and it could possibly form bridges between two HPV episomes or an episome and host DNA.

The striking co-localisation of CTCF and E2 binding sites outside the LCR of BPV1 leads to the speculation as to whether such co-localisation also exists in HPV. Using computer prediction and EMSA it would be possible to determine E2 binding sites outside the LCR. This would reveal if the pattern revealed in BPV1 also applies to HPV.

CTCF could be involved in blocking the estrogen receptor from acting on the estrogen response element what could explain how oestrogen promotes CIN progression. This

relationship is relatively easy to test with the cell lines generated in this study. The presence of an estrogen response element in HPV18 has to be confirmed and viral gene expression during oestrogen stimulation could be monitored in C3 mutant and wild type HPV18.

The PARlation status of CTCF is another important feature because PARlation of CTCF is needed for enhancer blocking, the protection from *de novo* methylation and CTCF-mediated PARlation of proteins in the vicinity of CTCF (Farrar *et al.*, 2010, Zampieri *et al.*, 2012). This could be done by ChIP using PAR antibody on samples of the cell lines generated in this study. However E2 binding sites outside of the LCR need to be determined first since PARlated E2 would also precipitate parts of the HPV genome and could lead to a false interpretation of the ChIP result if un-PARlated CTCF and PARlated E2 should co-localise. Furthermore a possible function of CTCF in promoting PARlation of E2 can be investigated by ChIP using E2 antibody and PAR antibody. The precipitated samples with E2 antibody would need to be precipitated again with PAR antibody check if E2 is PARlated. Using PAR antibody directly could give false positive results in case PARlated CTCF binds in the vicinity of E2. If an E2 binding site can be precipitated this way in the WT but not in the mutant it would mean that PARlation of E2 only occurs if CTCF binds to the HPV genome.

Also the expression of E1, E5, L1 and L2 in the C3 mutant is worth investigating. The viral helicase E1 can interfere with the tethering mechanism, so increased expression of E1 may be one possible reason for episome loss in the C3 mutant (Voitenleitner and Botchan, 2002). Furthermore the morphological abnormalities as well as the increased cell proliferation in the upper layers of the C3 mutant raft could be the result of increased E5 expression. This can be investigated by immunostaining or western blotting with functional E5 antibody. Also the expression of the late genes L1 and L2 may be altered in the C3 mutant if CTCF is involved in silencing these late genes. Again, this could be investigated by western blotting and immunostaining.

In conclusion, unravelling the details of how CTCF regulates HPV gene expression can improve our current understanding of HPV function and HPV-induced carcinogenesis. The enhanced understanding of the biological mechanisms in HPV-related disease has the potential to improve diagnostics as well as treatment and could eventually reduce the impact of this disease on the life of millions.

5 References

AKER, M., BOMSZTYK, K. & EMERY, D. W. 2010. Poly(ADP-ribose) polymerase-1 (PARP-1) contributes to the barrier function of a vertebrate chromatin insulator. *J Biol Chem,* 285**,** 37589-97.

AMELIO, A. L., MCANANY, P. K. & BLOOM, D. C. 2006. A chromatin insulator-like element in the herpes simplex virus type 1 latency-associated transcript region binds CCCTC-binding factor and displays enhancer-blocking and silencing activities. *J Virol,* 80**,** 2358-68.

ANGELETTI, P. C., KIM, K., FERNANDES, F. J. & LAMBERT, P. F. 2002. Stable replication of papillomavirus genomes in Saccharomyces cerevisiae. *J Virol,* 76**,** 3350-8.

ANTSON, A. A., BURNS, J. E., MOROZ, O. V., SCOTT, D. J., SANDERS, C. M., BRONSTEIN, I. B., DODSON, G. G., WILSON, K. S. & MAITLAND, N. J. 2000. Structure of the intact transactivation domain of the human papillomavirus E2 protein. *Nature,* 403**,** 805-9.

ARBEIT, J. M., HOWLEY, P. M. & HANAHAN, D. 1996. Chronic estrogen-induced cervical and vaginal squamous carcinogenesis in human papillomavirus type 16 transgenic mice. *Proc Natl Acad Sci U S A,* 93**,** 2930-5.

AU YEUNG, C. L., TSANG, W. P., TSANG, T. Y., CO, N. N., YAU, P. L. & KWOK, T. T. 2010. HPV-16 E6 upregulation of DNMT1 through repression of tumor suppressor p53. *Oncol Rep,* 24**,** 1599-604.

AVVAKUMOV, N., TORCHIA, J. & MYMRYK, J. S. 2003. Interaction of the HPV E7 proteins with the pCAF acetyltransferase. *Oncogene,* 22**,** 3833-41.

BADARACCO, G., VENUTI, A., SEDATI, A. & MARCANTE, M. L. 2002. HPV16 and HPV18 in genital tumors: Significantly different levels of viral integration and correlation to tumor invasiveness. *J Med Virol,* 67**,** 574-82.

BAKER, E. 2013. HPV and Pap: shifting roles in cervical cancer screening. *MLO Med Lab Obs,* 45**,** 24, 26.

BALDWIN, S. B., WALLACE, D. R., PAPENFUSS, M. R., ABRAHAMSEN, M., VAUGHT, L. C. & GIULIANO, A. R. 2004. Condom use and other factors affecting penile human papillomavirus detection in men attending a sexually transmitted disease clinic. *Sex Transm Dis,* 31**,** 601-7.

BANERJEE, N. S., WANG, H. K., BROKER, T. R. & CHOW, L. T. 2011. Human papillomavirus (HPV) E7 induces prolonged G2 following S phase reentry in differentiated human keratinocytes. *J Biol Chem,* 286**,** 15473-82.

BANERJEE, S., SMALLWOOD, A., LAMOND, S., CAMPBELL, S. & NARGUND, G. 2001. Igf2/H19 imprinting control region (ICR): an insulator or a position-dependent silencer? *ScientificWorldJournal,* 1**,** 218-24.

BANIAHMAD, A., STEINER, C., KOHNE, A. C. & RENKAWITZ, R. 1990. Modular structure of a chicken lysozyme silencer: involvement of an unusual thyroid hormone receptor binding site. *Cell,* 61**,** 505-14.

BAO, L., ZHOU, M. & CUI, Y. 2008. CTCFBSDB: a CTCF-binding site database for characterization of vertebrate genomic insulators. *Nucleic Acids Res,* 36**,** D83-7.

BARCELOS, A. C. & SOTTO, M. N. 2009. Comparative analysis of the expression of cytokeratins (1, 10, 14, 16, 4), involucrin, filaggrin and e-cadherin in plane warts and epidermodysplasia verruciformis plane wart-type lesions. *J Cutan Pathol,* 36**,** 647-54.

BARROW-LAING, L., CHEN, W. & ROMAN, A. 2010. Low- and high-risk human papillomavirus E7 proteins regulate p130 differently. *Virology,* 400**,** 233-9.

BARSKI, A., CUDDAPAH, S., CUI, K., ROH, T. Y., SCHONES, D. E., WANG, Z., WEI, G., CHEPELEV, I. & ZHAO, K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell,* 129**,** 823-37.

BASEMAN, J. G. & KOUTSKY, L. A. 2005. The epidemiology of human papillomavirus infections. *J Clin Virol,* 32 Suppl 1**,** S16-24.

BAYAS, J. M., COSTAS, L. & MUNOZ, A. 2008. Cervical cancer vaccination indications, efficacy, and side effects. *Gynecol Oncol,* 110**,** S11-4.

BEKKERS, R. L., MASSUGER, L. F., BULTEN, J. & MELCHERS, W. J. 2004. Epidemiological and clinical aspects of human papillomavirus detection in the prevention of cervical cancer. *Rev Med Virol,* 14**,** 95-105.

BELL, A. C. & FELSENFELD, G. 2000. Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature,* 405**,** 482-5.

BELL, A. C., WEST, A. G. & FELSENFELD, G. 1999. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell,* 98**,** 387-96.

BERGSTROM, R., WHITEHEAD, J., KURUKUTI, S. & OHLSSON, R. 2007. CTCF regulates asynchronous replication of the imprinted H19/Igf2 domain. *Cell Cycle,* 6**,** 450-4.

BERNARD, H. U. 2005. The clinical importance of the nomenclature, evolution and taxonomy of human papillomaviruses. *J Clin Virol,* 32 Suppl 1**,** S1-6.

BERNARD, H. U., BURK, R. D., CHEN, Z., VAN DOORSLAER, K., ZUR HAUSEN, H. & DE VILLIERS, E. M. 2010. Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments. *Virology,* 401**,** 70-9.

BISCHOF, O., NACERDDINE, K. & DEJEAN, A. 2005. Human papillomavirus oncoprotein E7 targets the promyelocytic leukemia protein and circumvents cellular senescence via the Rb and p53 tumor suppressor pathways. *Mol Cell Biol,* 25**,** 1013-24.

BLACK, P. H., HARTLEY, J. W., ROWE, W. P. & HUEBNER, R. J. 1963. Transformation of Bovine Tissue Culture Cells by Bovine Papilloma Virus. *Nature,* 199**,** 1016-8.

BLAKAJ, D. M., FERNANDEZ-FUENTES, N., CHEN, Z., HEGDE, R., FISER, A., BURK, R. D. & BRENOWITZ, M. 2009. Evolutionary and biophysical relationships among the papillomavirus E2 proteins. *Front Biosci,* 14**,** 900-17.

BLANTON, R. A., PEREZ-REYES, N., MERRICK, D. T. & MCDOUGALL, J. K. 1991. Epithelial cells immortalized by human papillomaviruses have premalignant characteristics in organotypic culture. *Am J Pathol,* 138**,** 673-85.

BLOOM, J., AMADOR, V., BARTOLINI, F., DEMARTINO, G. & PAGANO, M. 2003. Proteasome-mediated degradation of p21 via N-terminal ubiquitinylation. *Cell,* 115**,** 71-82.

BODILY, J. & LAIMINS, L. A. 2011. Persistence of human papillomavirus infection: keys to malignant progression. *Trends Microbiol,* 19**,** 33-9.

BOIRON, M., LEVY, J. P., THOMAS, M., FRIEDMANN, J. C. & BERNARD, J. 1964. Some Properties of Bovine Papilloma Virus. *Nature,* 201**,** 423-4.

BOLZER, A., KRETH, G., SOLOVEI, I., KOEHLER, D., SARACOGLU, K., FAUTH, C., MULLER, S., EILS, R., CREMER, C., SPEICHER, M. R. & CREMER, T. 2005. Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS Biol,* 3**,** e157.

BOSCH, F. X., BURCHELL, A. N., SCHIFFMAN, M., GIULIANO, A. R., DE SANJOSE, S., BRUNI, L., TORTOLERO-LUNA, G., KJAER, S. K. & MUNOZ, N. 2008. Epidemiology and natural history of human papillomavirus infections and type-specific implications in cervical neoplasia. *Vaccine,* 26 Suppl 10**,** K1-16.

BOSCH, F. X. & DE SANJOSE, S. 2003. Chapter 1: Human papillomavirus and cervical cancer--burden and assessment of causality. *J Natl Cancer Inst Monogr***,** 3-13.

BOTTA, M., HAIDER, S., LEUNG, I. X., LIO, P. & MOZZICONACCI, J. 2010. Intra- and inter-chromosomal interactions correlate with CTCF binding genome wide. *Mol Syst Biol,* 6**,** 426.

BOTTALICO, D., CHEN, Z., DUNNE, A., OSTOLOZA, J., MCKINNEY, S., SUN, C., SCHLECHT, N. F., FATAHZADEH, M., HERRERO, R., SCHIFFMAN, M. & BURK, R. D. 2011. The oral cavity contains abundant known and novel human papillomaviruses from the Betapapillomavirus and Gammapapillomavirus genera. *J Infect Dis,* 204**,** 787-92.

BOUVARD, V., BAAN, R., STRAIF, K., GROSSE, Y., SECRETAN, B., EL GHISSASSI, F., BENBRAHIM-TALLAA, L., GUHA, N., FREEMAN, C., GALICHET, L. & COGLIANO, V. 2009. A review of human carcinogens--Part B: biological agents. *Lancet Oncol,* 10**,** 321-2.

BRUNI, L., DIAZ, M., CASTELLSAGUE, X., FERRER, E., BOSCH, F. X. & DE SANJOSE, S. 2010. Cervical human papillomavirus prevalence in 5 continents: meta-analysis of 1 million women with normal cytological findings. *J Infect Dis,* 202**,** 1789-99.

BUCK, C. B., CHENG, N., THOMPSON, C. D., LOWY, D. R., STEVEN, A. C., SCHILLER, J. T. & TRUS, B. L. 2008. Arrangement of L2 within the papillomavirus capsid. *J Virol,* 82**,** 5190-7.

BUCK, C. B., THOMPSON, C. D., PANG, Y. Y., LOWY, D. R. & SCHILLER, J. T. 2005. Maturation of papillomavirus capsids. *J Virol,* 79**,** 2839-46.

BURCIN, M., ARNOLD, R., LUTZ, M., KAISER, B., RUNGE, D., LOTTSPEICH, F., FILIPPOVA, G. N., LOBANENKOV, V. V. & RENKAWITZ, R. 1997. Negative protein 1, which is required for function of the chicken lysozyme gene silencer in conjunction with hormone receptors, is identical to the multivalent zinc finger repressor CTCF. *Mol Cell Biol,* 17**,** 1281-8.

CAIAFA, P., GUASTAFIERRO, T. & ZAMPIERI, M. 2009. Epigenetics: poly(ADP-ribosyl)ation of PARP-1 regulates genomic methylation patterns. *FASEB J,* 23**,** 672-8.

CAMPO, M. S., GRAHAM, S. V., CORTESE, M. S., ASHRAFI, G. H., ARAIBI, E. H., DORNAN, E. S., MINERS, K., NUNES, C. & MAN, S. 2010. HPV-16 E5 down-regulates expression of surface HLA class I and reduces recognition by CD8 T cells. *Virology,* 407**,** 137-42.

CASTLE, P. E. 2004. Beyond human papillomavirus: the cervix, exogenous secondary factors, and the development of cervical precancer and cancer. *J Low Genit Tract Dis,* 8**,** 224-30.

CHAIWONGKOT, A., VINOKUROVA, S., PIENTONG, C., EKALAKSANANAN, T., KONGYINGYOES, B., KLEEBKAOW, P., CHUMWORATHAYI, B., PATARAPADUNGKIT, N., REUSCHENBACH, M. & VON KNEBEL DOEBERITZ, M. 2013. Differential methylation of E2 binding sites in episomal and integrated HPV 16 genomes in preinvasive and invasive cervical lesions. *Int J Cancer,* 132**,** 2087-94.

CHAU, C. M., ZHANG, X. Y., MCMAHON, S. B. & LIEBERMAN, P. M. 2006. Regulation of Epstein-Barr virus latency type by the chromatin boundary factor CTCF. *J Virol,* 80**,** 5723-32.

CHEN, H., TIAN, Y., SHU, W., BO, X. & WANG, S. 2012a. Comprehensive identification and annotation of cell type-specific and ubiquitous CTCF-binding sites in the human genome. *PLoS One,* 7**,** e41374.

CHEN, H. S., WIKRAMASINGHE, P., SHOWE, L. & LIEBERMAN, P. M. 2012b. Cohesins repress Kaposi's sarcoma-associated herpesvirus immediate early gene transcription during latency. *J Virol,* 86**,** 9454-64.

CHERNUKHIN, I., SHAMSUDDIN, S., KANG, S. Y., BERGSTROM, R., KWON, Y. W., YU, W., WHITEHEAD, J., MUKHOPADHYAY, R., DOCQUIER, F., FARRAR, D., MORRISON, I., VIGNERON, M., WU, S. Y., CHIANG, C. M., LOUKINOV, D., LOBANENKOV, V., OHLSSON, R. & KLENOVA, E. 2007. CTCF interacts with and recruits the largest

subunit of RNA polymerase II to CTCF target sites genome-wide. *Mol Cell Biol,* 27**,** 1631-48.

CHEUNG, J. L., CHEUNG, T. H., YU, M. Y. & CHAN, P. K. 2013. Virological characteristics of cervical cancers carrying pure episomal form of HPV16 genome. *Gynecol Oncol*. 131, 374-9.

CHIEN, R., ZENG, W., KAWAUCHI, S., BENDER, M. A., SANTOS, R., GREGSON, H. C., SCHMIESING, J. A., NEWKIRK, D. A., KONG, X., BALL, A. R., JR., CALOF, A. L., LANDER, A. D., GROUDINE, M. T. & YOKOMORI, K. 2011. Cohesin mediates chromatin interactions that regulate mammalian beta-globin expression. *J Biol Chem,* 286**,** 17870-8.

CLIFFORD, G. M., SMITH, J. S., PLUMMER, M., MUNOZ, N. & FRANCESCHI, S. 2003. Human papillomavirus types in invasive cervical cancer worldwide: a meta-analysis. *Br J Cancer,* 88**,** 63-73.

COMET, I., SCHUETTENGRUBER, B., SEXTON, T. & CAVALLI, G. 2011. A chromatin insulator driving three-dimensional Polycomb response element (PRE) contacts and Polycomb association with the chromatin fiber. *Proc Natl Acad Sci U S A,* 108**,** 2294-9.

CONE, R. W., MINSON, A. C., SMITH, M. R. & MCDOUGALL, J. K. 1992. Conservation of HPV-16 E6/E7 ORF sequences in a cervical carcinoma. *J Med Virol,* 37**,** 99-107.

COTTLER, L., GARVIN, E. C. & CALLAHAN, C. 2006. Condom use and the risk of HPV infection. *N Engl J Med,* 355**,** 1388-9; author reply 1389.

CREMER, T. & CREMER, C. 2001. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat Rev Genet,* 2**,** 292-301.

CROSBIE, E. J. & KITCHENER, H. C. 2007. Cervarix--a bivalent L1 virus-like particle vaccine for prevention of human papillomavirus type 16- and 18-associated cervical cancer. *Expert Opin Biol Ther,* 7**,** 391-6.

CRUSIUS, K., AUVINEN, E., STEUER, B., GAISSERT, H. & ALONSO, A. 1998. The human papillomavirus type 16 E5-protein modulates ligand-dependent activation of the EGF receptor family in the human epithelial cell line HaCaT. *Exp Cell Res,* 241**,** 76-83.

CRUSIUS, K., RODRIGUEZ, I. & ALONSO, A. 2000. The human papillomavirus type 16 E5 protein modulates ERK1/2 and p38 MAP kinase activation by an EGFR-independent process in stressed human keratinocytes. *Virus Genes,* 20**,** 65-9.

CUTTS, F. T., FRANCESCHI, S., GOLDIE, S., CASTELLSAGUE, X., DE SANJOSE, S., GARNETT, G., EDMUNDS, W. J., CLAEYS, P., GOLDENTHAL, K. L., HARPER, D. M. & MARKOWITZ, L. 2007. Human papillomavirus and HPV vaccines: a review. *Bull World Health Organ,* 85**,** 719-26.

DALL, K. L., SCARPINI, C. G., ROBERTS, I., WINDER, D. M., STANLEY, M. A., MURALIDHAR, B., HERDMAN, M. T., PETT, M. R. & COLEMAN, N. 2008. Characterization of naturally occurring HPV16 integration sites isolated from cervical keratinocytes under noncompetitive conditions. *Cancer Res,* 68**,** 8249-59.

DAO, L. D., DUFFY, A., VAN TINE, B. A., WU, S. Y., CHIANG, C. M., BROKER, T. R. & CHOW, L. T. 2006. Dynamic localization of the human papillomavirus type 11 origin binding protein E2 through mitosis while in association with the spindle apparatus. *J Virol,* 80**,** 4792-800.

DARNELL, G. A., SCHRODER, W. A., ANTALIS, T. M., LAMBLEY, E., MAJOR, L., GARDNER, J., BIRRELL, G., CID-ARREGUI, A. & SUHRBIER, A. 2007. Human papillomavirus E7 requires the protease calpain to degrade the retinoblastoma protein. *J Biol Chem,* 282**,** 37492-500.

DAVALOS-SALAS, M., FURLAN-MAGARIL, M., GONZALEZ-BUENDIA, E., VALDES-QUEZADA, C., AYALA-ORTEGA, E. & RECILLAS-TARGA, F. 2011. Gain of DNA methylation is enhanced in the absence of CTCF at the human retinoblastoma gene promoter. *BMC Cancer,* 11**,** 232.

DAVY, C. E., AYUB, M., JACKSON, D. J., DAS, P., MCINTOSH, P. & DOORBAR, J. 2006. HPV16 E1--E4 protein is phosphorylated by Cdk2/cyclin A and relocalizes this complex to the cytoplasm. *Virology,* 349**,** 230-44.

DAVY, C. E., JACKSON, D. J., WANG, Q., RAJ, K., MASTERSON, P. J., FENNER, N. F., SOUTHERN, S., CUTHILL, S., MILLAR, J. B. & DOORBAR, J. 2002. Identification of a G(2) arrest domain in the E1 wedge E4 protein of human papillomavirus type 16. *J Virol,* 76**,** 9806-18.

DAY, L., CHAU, C. M., NEBOZHYN, M., RENNEKAMP, A. J., SHOWE, M. & LIEBERMAN, P. M. 2007. Chromatin profiling of Epstein-Barr virus latency control region. *J Virol,* 81**,** 6389-401.

DAY, P. M., RODEN, R. B., LOWY, D. R. & SCHILLER, J. T. 1998. The papillomavirus minor capsid protein, L2, induces localization of the major capsid protein, L1, and the viral transcription/replication protein, E2, to PML oncogenic domains. *J Virol,* 72**,** 142-50.

DE-CASTRO ARCE, J., GOCKEL-KRZIKALLA, E. & ROSL, F. 2012. Silencing of multi-copy HPV16 by viral self-methylation and chromatin occlusion: a model for epigenetic virus-host interaction. *Hum Mol Genet,* 21**,** 1693-705.

DE MARTEL, C., FERLAY, J., FRANCESCHI, S., VIGNAT, J., BRAY, F., FORMAN, D. & PLUMMER, M. 2012. Global burden of cancers attributable to infections in 2008: a review and synthetic analysis. *Lancet Oncol,* 13**,** 607-15.

DE SANJOSE, S., QUINT, W. G., ALEMANY, L., GERAETS, D. T., KLAUSTERMEIER, J. E., LLOVERAS, B., TOUS, S., FELIX, A., BRAVO, L. E., SHIN, H. R., VALLEJOS, C. S., DE RUIZ, P. A., LIMA, M. A., GUIMERA, N., CLAVERO, O., ALEJO, M., LLOMBART-BOSCH, A., CHENG-YANG, C., TATTI, S. A., KASAMATSU, E., ILJAZOVIC, E., ODIDA, M., PRADO, R., SEOUD, M., GRCE, M., USUBUTUN, A., JAIN, A., SUAREZ, G. A., LOMBARDI, L. E., BANJO, A., MENENDEZ, C., DOMINGO, E. J., VELASCO, J., NESSA, A., CHICHAREON, S. C., QIAO, Y. L., LERMA, E., GARLAND, S. M., SASAGAWA, T., FERRERA, A., HAMMOUDA, D., MARIANI, L., PELAYO, A., STEINER, I., OLIVA, E., MEIJER, C. J., AL-JASSAR, W. F., CRUZ, E., WRIGHT, T. C., PURAS, A., LLAVE, C. L., TZARDI, M., AGORASTOS, T., GARCIA-BARRIOLA, V., CLAVEL, C., ORDI, J., ANDUJAR, M., CASTELLSAGUE, X., SANCHEZ, G. I., NOWAKOWSKI, A. M., BORNSTEIN, J., MUNOZ, N. & BOSCH, F. X. 2010. Human papillomavirus genotype attribution in invasive cervical cancer: a retrospective cross-sectional worldwide study. *Lancet Oncol,* 11**,** 1048-56.

DELL, G., WILKINSON, K. W., TRANTER, R., PARISH, J., LEO BRADY, R. & GASTON, K. 2003. Comparison of the structure and DNA-binding properties of the E2 proteins from an oncogenic and a non-oncogenic human papillomavirus. *J Mol Biol,* 334**,** 979-91.

DEMASI, J., HUH, K. W., NAKATANI, Y., MUNGER, K. & HOWLEY, P. M. 2005. Bovine papillomavirus E7 transformation function correlates with cellular p600 protein binding. *Proc Natl Acad Sci U S A,* 102**,** 11486-91.

DERKAY, C. S. 1995. Task force on recurrent respiratory papillomas. A preliminary report. *Arch Otolaryngol Head Neck Surg,* 121**,** 1386-91.

DESAI, S., WETTEN, S., WOODHALL, S. C., PETERS, L., HUGHES, G. & SOLDAN, K. 2011. Genital warts and cost of care in England. *Sex Transm Infect,* 87**,** 464-8.

DONNE, A. J. & CLARKE, R. 2010. Recurrent respiratory papillomatosis: an uncommon but potentially devastating effect of human papillomavirus in children. *Int J STD AIDS,* 21**,** 381-5.

DOORBAR, J. 2005. The papillomavirus life cycle. *J Clin Virol,* 32 Suppl 1**,** S7-15.

DOORBAR, J. 2006. Molecular biology of human papillomavirus infection and cervical cancer. *Clin Sci (Lond),* 110**,** 525-41.

DOORBAR, J. 2013. Latent papillomavirus infections and their regulation. *Curr Opin Virol,* 3**,** 416-21.

DOORBAR, J., FOO, C., COLEMAN, N., MEDCALF, L., HARTLEY, O., PROSPERO, T., NAPTHINE, S., STERLING, J., WINTER, G. & GRIFFIN, H. 1997. Characterization of events during the late stages of HPV16 infection in vivo using high-affinity synthetic Fabs to E4. *Virology,* 238**,** 40-52.

DOORBAR, J., QUINT, W., BANKS, L., BRAVO, I. G., STOLER, M., BROKER, T. R. & STANLEY, M. A. 2012. The biology and life-cycle of human papillomaviruses. *Vaccine,* 30 Suppl 5**,** F55-70.

DOSTATNI, N., LAMBERT, P. F., SOUSA, R., HAM, J., HOWLEY, P. M. & YANIV, M. 1991. The functional BPV-1 E2 trans-activating protein can act as a repressor by preventing formation of the initiation complex. *Genes Dev,* 5**,** 1657-71.

DUNN, K. L., ZHAO, H. & DAVIE, J. R. 2003. The insulator binding protein CTCF associates with the nuclear matrix. *Exp Cell Res,* 288**,** 218-23.

EDWARDS, T. G., KOELLER, K. J., SLOMCZYNSKA, U., FOK, K., HELMUS, M., BASHKIN, J. K. & FISHER, C. 2011. HPV episome levels are potently decreased by pyrrole-imidazole polyamides. *Antiviral Res,* 91**,** 177-86.

EGAWA, N., NAKAHARA, T., OHNO, S., NARISAWA-SAITO, M., YUGAWA, T., FUJITA, M., YAMATO, K., NATORI, Y. & KIYONO, T. 2012. The E1 protein of human papillomavirus type 16 is dispensable for maintenance replication of the viral genome. *J Virol,* 86**,** 3276-83.

EKSTROM, J., BZHALAVA, D., SVENBACK, D., FORSLUND, O. & DILLNER, J. 2011. High throughput sequencing reveals diversity of Human Papillomaviruses in cutaneous lesions. *Int J Cancer,* 129**,** 2643-50.

EKSTROM, J., FORSLUND, O. & DILLNER, J. 2010. Three novel papillomaviruses (HPV109, HPV112 and HPV114) and their presence in cutaneous and mucosal samples. *Virology,* 397**,** 331-6.

EL-KADY, A. & KLENOVA, E. 2005. Regulation of the transcription factor, CTCF, by phosphorylation with protein kinase CK2. *FEBS Lett,* 579**,** 1424-34.

ERTEL, M. K., CAMMARATA, A. L., HRON, R. J. & NEUMANN, D. M. 2012. CTCF occupation of the herpes simplex virus 1 genome is disrupted at early times postreactivation in a transcription-dependent manner. *J Virol,* 86**,** 12741-59.

ESSIEN, K., VIGNEAU, S., APRELEVA, S., SINGH, L. N., BARTOLOMEI, M. S. & HANNENHALLI, S. 2009. CTCF binding site classes exhibit distinct evolutionary, genomic, epigenomic and transcriptomic features. *Genome Biol,* 10**,** R131.

FARRAR, D., RAI, S., CHERNUKHIN, I., JAGODIC, M., ITO, Y., YAMMINE, S., OHLSSON, R., MURRELL, A. & KLENOVA, E. 2010. Mutational analysis of the poly(ADP-ribosyl)ation sites of the transcription factor CTCF provides an insight into the mechanism of its regulation by poly(ADP-ribosyl)ation. *Mol Cell Biol,* 30**,** 1199-216.

FEDORIW, A. M., STEIN, P., SVOBODA, P., SCHULTZ, R. M. & BARTOLOMEI, M. S. 2004. Transgenic RNAi reveals essential function for CTCF in H19 gene imprinting. *Science,* 303**,** 238-40.

FELSANI, A., MILEO, A. M. & PAGGI, M. G. 2006. Retinoblastoma family proteins as key targets of the small DNA virus oncoproteins. *Oncogene,* 25**,** 5277-85.

FILIPPOVA, G. N., FAGERLIE, S., KLENOVA, E. M., MYERS, C., DEHNER, Y., GOODWIN, G., NEIMAN, P. E., COLLINS, S. J. & LOBANENKOV, V. V. 1996. An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc

fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. *Mol Cell Biol,* 16**,** 2802-13.

FINNEN, R. L., ERICKSON, K. D., CHEN, X. S. & GARCEA, R. L. 2003. Interactions between papillomavirus L1 and L2 capsid proteins. *J Virol,* 77**,** 4818-26.

FLICK, K., OUNI, I., WOHLSCHLEGEL, J. A., CAPATI, C., MCDONALD, W. H., YATES, J. R. & KAISER, P. 2004. Proteolysis-independent regulation of the transcription factor Met4 by a single Lys 48-linked ubiquitin chain. *Nat Cell Biol,* 6**,** 634-41.

FLORIN, L., SAPP, C., STREECK, R. E. & SAPP, M. 2002. Assembly and translocation of papillomavirus capsid proteins. *J Virol,* 76**,** 10009-14.

FORMAN, D., DE MARTEL, C., LACEY, C. J., SOERJOMATARAM, I., LORTET-TIEULENT, J., BRUNI, L., VIGNAT, J., FERLAY, J., BRAY, F., PLUMMER, M. & FRANCESCHI, S. 2012. Global burden of human papillomavirus and related diseases. *Vaccine,* 30 Suppl 5**,** F12-23.

FORSLUND, O. 2007. Genetic diversity of cutaneous human papillomaviruses. *J Gen Virol,* 88**,** 2662-9.

FRANCESCHI, S., HERRERO, R., CLIFFORD, G. M., SNIJDERS, P. J., ARSLAN, A., ANH, P. T., BOSCH, F. X., FERRECCIO, C., HIEU, N. T., LAZCANO-PONCE, E., MATOS, E., MOLANO, M., QIAO, Y. L., RAJKUMAR, R., RONCO, G., DE SANJOSE, S., SHIN, H. R., SUKVIRACH, S., THOMAS, J. O., MEIJER, C. J. & MUNOZ, N. 2006. Variations in the age-specific curves of human papillomavirus prevalence in women worldwide. *Int J Cancer,* 119**,** 2677-84.

FRATTINI, M. G., LIM, H. B., DOORBAR, J. & LAIMINS, L. A. 1997. Induction of human papillomavirus type 18 late gene expression and genomic amplification in organotypic cultures from transfected DNA templates. *J Virol,* 71**,** 7068-72.

FU, L., VAN DOORSLAER, K., CHEN, Z., RISTRIANI, T., MASSON, M., TRAVE, G. & BURK, R. D. 2010. Degradation of p53 by human Alphapapillomavirus E6 proteins shows a stronger correlation with phylogeny than oncogenicity. *PLoS One,* 5, e12816.

GAMMOH, N., GRM, H. S., MASSIMI, P. & BANKS, L. 2006. Regulation of human papillomavirus type 16 E7 activity through direct protein interaction with the E2 transcriptional activator. *J Virol,* 80**,** 1787-97.

GARCIA-VALLVE, S., ALONSO, A. & BRAVO, I. G. 2005. Papillomaviruses: different genes have different histories. *Trends Microbiol,* 13**,** 514-21.

GARIGLIO, P., GUTIERREZ, J., CORTES, E. & VAZQUEZ, J. 2009. The role of retinoid deficiency and estrogens as cofactors in cervical cancer. *Arch Med Res,* 40**,** 449-65.

GERAETS, D., ALEMANY, L., GUIMERA, N., DE SANJOSE, S., DE KONING, M., MOLIJN, A., JENKINS, D., BOSCH, X. & QUINT, W. 2012. Detection of rare and possibly carcinogenic human papillomavirus genotypes as single infections in invasive cervical cancer. *J Pathol*, 10, 1002

GEREIN, V., RASTORGUEV, E., GEREIN, J., DRAF, W. & SCHIRREN, J. 2005. Incidence, age at onset, and potential reasons of malignant transformation in recurrent respiratory papillomatosis patients: 20 years experience. *Otolaryngol Head Neck Surg,* 132**,** 392-4.

GEWIN, L., MYERS, H., KIYONO, T. & GALLOWAY, D. A. 2004. Identification of a novel telomerase repressor that interacts with the human papillomavirus type-16 E6/E6-AP complex. *Genes Dev,* 18**,** 2269-82.

GILLISON, M. L., BROUTIAN, T., PICKARD, R. K., TONG, Z. Y., XIAO, W., KAHLE, L., GRAUBARD, B. I. & CHATURVEDI, A. K. 2012. Prevalence of oral HPV infection in the United States, 2009-2010. *JAMA,* 307**,** 693-703.

GLICKMAN, M. H. & CIECHANOVER, A. 2002. The ubiquitin-proteasome proteolytic pathway: destruction for the sake of construction. *Physiol Rev,* 82**,** 373-428.

GRAHAM, S. V. 2010. Human papillomavirus: gene expression, regulation and prospects for novel diagnostic methods and antiviral therapies. *Future Microbiol,* 5**,** 1493-506.

GRAY, E., PETT, M. R., WARD, D., WINDER, D. M., STANLEY, M. A., ROBERTS, I., SCARPINI, C. G. & COLEMAN, N. 2010. In vitro progression of human papillomavirus 16 episome-associated cervical neoplasia displays fundamental similarities to integrant-associated carcinogenesis. *Cancer Res,* 70**,** 4081-91.

GREEN, H. 1977. Terminal differentiation of cultured human epidermal cells. *Cell,* 11**,** 405-16.

GRIFFIN, H., WU, Z., MARNANE, R., DEWAR, V., MOLIJN, A., QUINT, W., VAN HOOF, C., STRUYF, F., COLAU, B., JENKINS, D. & DOORBAR, J. 2012. E4 antibodies facilitate detection and type-assignment of active HPV infection in cervical disease. *PLoS One,* 7**,** e49974.

GRM, H. S., MASSIMI, P., GAMMOH, N. & BANKS, L. 2005. Crosstalk between the human papillomavirus E2 transcriptional activator and the E6 oncoprotein. *Oncogene,* 24**,** 5149-64.

GSCHWANDTNER, M., MILDNER, M., MLITZ, V., GRUBER, F., ECKHART, L., WERFEL, T., GUTZMER, R., ELIAS, P. M. & TSCHACHLER, E. 2013. Histamine suppresses epidermal keratinocyte differentiation and impairs skin barrier function in a human skin model. *Allergy,* 68**,** 37-47.

GUASTAFIERRO, T., CECCHINELLI, B., ZAMPIERI, M., REALE, A., RIGGIO, G., STHANDIER, O., ZUPI, G., CALABRESE, L. & CAIAFA, P. 2008. CCCTC-binding factor activates PARP-1 affecting DNA methylation machinery. *J Biol Chem,* 283**,** 21873-80.

GUELEN, L., PAGIE, L., BRASSET, E., MEULEMAN, W., FAZA, M. B., TALHOUT, W., EUSSEN, B. H., DE KLEIN, A., WESSELS, L., DE LAAT, W. & VAN STEENSEL, B. 2008. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature,* 453**,** 948-51.

HADJUR, S., WILLIAMS, L. M., RYAN, N. K., COBB, B. S., SEXTON, T., FRASER, P., FISHER, A. G. & MERKENSCHLAGER, M. 2009. Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus. *Nature,* 460**,** 410-3.

HAGA, K., OHNO, S., YUGAWA, T., NARISAWA-SAITO, M., FUJITA, M., SAKAMOTO, M., GALLOWAY, D. A. & KIYONO, T. 2007. Efficient immortalization of primary human cells by p16INK4a-specific short hairpin RNA or Bmi-1, combined with introduction of hTERT. *Cancer Sci,* 98**,** 147-54.

HANDOKO, L., XU, H., LI, G., NGAN, C. Y., CHEW, E., SCHNAPP, M., LEE, C. W., YE, C., PING, J. L., MULAWADI, F., WONG, E., SHENG, J., ZHANG, Y., POH, T., CHAN, C. S., KUNARSO, G., SHAHAB, A., BOURQUE, G., CACHEUX-RATABOUL, V., SUNG, W. K., RUAN, Y. & WEI, C. L. 2011. CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat Genet,* 43**,** 630-8.

HARK, A. T., SCHOENHERR, C. J., KATZ, D. J., INGRAM, R. S., LEVORSE, J. M. & TILGHMAN, S. M. 2000. CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature,* 405**,** 486-9.

HARPER, D. M., FRANCO, E. L., WHEELER, C. M., MOSCICKI, A. B., ROMANOWSKI, B., ROTELI-MARTINS, C. M., JENKINS, D., SCHUIND, A., COSTA CLEMENS, S. A. & DUBIN, G. 2006. Sustained efficacy up to 4.5 years of a bivalent L1 virus-like particle vaccine against human papillomavirus types 16 and 18: follow-up from a randomised control trial. *Lancet,* 367**,** 1247-55.

HARTLEY, K. A. & ALEXANDER, K. A. 2002. Human TATA binding protein inhibits human papillomavirus type 11 DNA replication by antagonizing E1-E2 protein complex formation on the viral origin of replication. *J Virol,* 76**,** 5014-23.

HAVERKOS, H. W. 2004. Viruses, chemicals and co-carcinogenesis. *Oncogene,* 23**,** 6492-9.

HEATH, H., RIBEIRO DE ALMEIDA, C., SLEUTELS, F., DINGJAN, G., VAN DE NOBELEN, S., JONKERS, I., LING, K. W., GRIBNAU, J., RENKAWITZ, R., GROSVELD, F., HENDRIKS, R. W. & GALJART, N. 2008. CTCF regulates cell cycle progression of alphabeta T cells in the thymus. *EMBO J,* 27**,** 2839-50.

HERFS, M., YAMAMOTO, Y., LAURY, A., WANG, X., NUCCI, M. R., MCLAUGHLIN-DRUBIN, M. E., MUNGER, K., FELDMAN, S., MCKEON, F. D., XIAN, W. & CRUM, C. P. 2012. A discrete population of squamocolumnar junction cells implicated in the pathogenesis of cervical cancer. *Proc Natl Acad Sci U S A,* 109**,** 10516-21.

HERNANDEZ-RAMON, E. E., BURNS, J. E., ZHANG, W., WALKER, H. F., ALLEN, S., ANTSON, A. A. & MAITLAND, N. J. 2008. Dimerization of the human papillomavirus type 16 E2 N terminus results in DNA looping within the upstream regulatory region. *J Virol,* 82**,** 4853-61.

HERRINGTON, C. S. 2009. Recent advances in molecular gynaecological pathology. *Histopathology,* 55**,** 243-9.

HIBMA, M. H. 2012. The immune response to papillomavirus during infection persistence and regression. *Open Virol J,* 6**,** 241-8.

HOFFMANN, R., HIRT, B., BECHTOLD, V., BEARD, P. & RAJ, K. 2006. Different modes of human papillomavirus DNA replication during maintenance. *J Virol,* 80**,** 4431-9.

HOHL, D., MEHREL, T., LICHTI, U., TURNER, M. L., ROOP, D. R. & STEINERT, P. M. 1991. Characterization of human loricrin. Structure and function of a new class of epidermal cell envelope proteins. *J Biol Chem,* 266**,** 6626-36.

HOLMGREN, S. C., PATTERSON, N. A., OZBUN, M. A. & LAMBERT, P. F. 2005. The minor capsid protein L2 contributes to two steps in the human papillomavirus type 31 life cycle. *J Virol,* 79**,** 3938-48.

HOLWERDA, S. J. & DE LAAT, W. 2013. CTCF: the protein, the binding partners, the binding sites and their chromatin loops. *Philos Trans R Soc Lond B Biol Sci,* 368**,** 20120369.

HOU, S. Y., WU, S. Y. & CHIANG, C. M. 2002. Transcriptional activity among high and low risk human papillomavirus E2 proteins correlates with E2 DNA binding. *J Biol Chem,* 277**,** 45619-29.

HSUEH, P. R. 2009. Human papillomavirus, genital warts, and vaccines. *J Microbiol Immunol Infect,* 42**,** 101-6.

HUBBERT, N. L., SCHILLER, J. T., LOWY, D. R. & ANDROPHY, E. J. 1988. Bovine papilloma virus-transformed cells contain multiple E2 proteins. *Proc Natl Acad Sci U S A,* 85**,** 5864-8.

HUBERT, W. G. & LAIMINS, L. A. 2002. Human papillomavirus type 31 replication modes during the early phases of the viral life cycle depend on transcriptional and posttranscriptional regulation of E1 and E2 expression. *J Virol,* 76**,** 2263-73.

HUH, K. W., DEMASI, J., OGAWA, H., NAKATANI, Y., HOWLEY, P. M. & MUNGER, K. 2005. Association of the human papillomavirus type 16 E7 oncoprotein with the 600-kDa retinoblastoma protein-associated factor, p600. *Proc Natl Acad Sci U S A,* 102**,** 11492-7.

HYLAND, P. L., MCDADE, S. S., MCCLOSKEY, R., DICKSON, G. J., ARTHUR, K., MCCANCE, D. J. & PATEL, D. 2011. Evidence for alteration of EZH2, BMI1, and KDM6A and epigenetic reprogramming in human papillomavirus type 16 E6/E7-expressing keratinocytes. *J Virol,* 85**,** 10999-1006.

ILVES, I., KADAJA, M. & USTAV, M. 2003. Two separate replication modes of the bovine papillomavirus BPV1 origin of replication that have different sensitivity to p53. *Virus Res,* 96**,** 75-84.

JACKSON, S. & STOREY, A. 2000. E6 proteins from diverse cutaneous HPV types inhibit apoptosis in response to UV damage. *Oncogene,* 19**,** 592-8.

JEON, S., ALLEN-HOFFMANN, B. L. & LAMBERT, P. F. 1995. Integration of human papillomavirus type 16 into the human genome correlates with a selective growth advantage of cells. *J Virol,* 69**,** 2989-97.

JEON, S. & LAMBERT, P. F. 1995. Integration of human papillomavirus type 16 DNA into the human genome leads to increased stability of E6 and E7 mRNAs: implications for cervical carcinogenesis. *Proc Natl Acad Sci U S A,* 92**,** 1654-8.

JOHANNSEN, E. & LAMBERT, P. F. 2013. Epigenetics of human papillomaviruses. *Virology,* 445**,** 205-12.

JOHANSSON, C., SOMBERG, M., LI, X., BACKSTROM WINQUIST, E., FAY, J., RYAN, F., PIM, D., BANKS, L. & SCHWARTZ, S. 2012. HPV-16 E2 contributes to induction of HPV-16 late gene expression by inhibiting early polyadenylation. *EMBO J,* 31**,** 3212-27.

KABSCH, K., MOSSADEGH, N., KOHL, A., KOMPOSCH, G., SCHENKEL, J., ALONSO, A. & TOMAKIDI, P. 2004. The HPV-16 E5 protein inhibits TRAIL- and FasL-mediated apoptosis in human keratinocyte raft cultures. *Intervirology,* 47**,** 48-56.

KANDA, T., KAMIYA, M., MARUO, S., IWAKIRI, D. & TAKADA, K. 2007. Symmetrical localization of extrachromosomally replicating viral genomes on sister chromatids. *J Cell Sci,* 120**,** 1529-39.

KANDURI, C., PANT, V., LOUKINOV, D., PUGACHEVA, E., QI, C. F., WOLFFE, A., OHLSSON, R. & LOBANENKOV, V. V. 2000. Functional association of CTCF with the insulator upstream of the H19 gene is parent of origin-specific and methylation-sensitive. *Curr Biol,* 10**,** 853-6.

KANDURI, M., KANDURI, C., MARIANO, P., VOSTROV, A. A., QUITSCHKE, W., LOBANENKOV, V. & OHLSSON, R. 2002. Multiple nucleosome positioning sites regulate the CTCF-mediated insulator function of the H19 imprinting control region. *Mol Cell Biol,* 22**,** 3339-44.

KANG, H., CHO, H., SUNG, G. H. & LIEBERMAN, P. M. 2013. CTCF regulates Kaposi's sarcoma-associated herpesvirus latency transcription by nucleosome displacement and RNA polymerase programming. *J Virol,* 87**,** 1789-99.

KANODIA, S., FAHEY, L. M. & KAST, W. M. 2007. Mechanisms used by human papillomaviruses to escape the host immune response. *Curr Cancer Drug Targets,* 7**,** 79-89.

KARIM, R., MEYERS, C., BACKENDORF, C., LUDIGS, K., OFFRINGA, R., VAN OMMEN, G. J., MELIEF, C. J., VAN DER BURG, S. H. & BOER, J. M. 2011. Human papillomavirus deregulates the response of a cellular network comprising of chemotactic and proinflammatory genes. *PLoS One,* 6**,** e17848.

KEL, A. E., GOSSLING, E., REUTER, I., CHEREMUSHKIN, E., KEL-MARGOULIS, O. V. & WINGENDER, E. 2003. MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res,* 31**,** 3576-9.

KIM, K., GARNER-HAMRICK, P. A., FISHER, C., LEE, D. & LAMBERT, P. F. 2003. Methylation patterns of papillomavirus DNA, its influence on E2 function, and implications in viral infection. *J Virol,* 77**,** 12450-9.

KIM, K. & LAMBERT, P. F. 2002. E1 protein of bovine papillomavirus 1 is not required for the maintenance of viral plasmid DNA replication. *Virology,* 293**,** 10-4.

KIM, T. H., ABDULLAEV, Z. K., SMITH, A. D., CHING, K. A., LOUKINOV, D. I., GREEN, R. D., ZHANG, M. Q., LOBANENKOV, V. V. & REN, B. 2007. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell,* 128**,** 1231-45.

KITASATO, H., TARDY-KADLEC, A., SMAHEL, M., TYKVA, R. & VONKA, V. 1997. Putative regulatory sequence in human papillomavirus type 16 E2 open reading frame. *Folia Biol (Praha),* 43**,** 41-4.

KIYONO, T., FOSTER, S. A., KOOP, J. I., MCDOUGALL, J. K., GALLOWAY, D. A. & KLINGELHUTZ, A. J. 1998. Both Rb/p16INK4a inactivation and telomerase activity are required to immortalize human epithelial cells. *Nature,* 396**,** 84-8.

KIYONO, T., HIRAIWA, A., FUJITA, M., HAYASHI, Y., AKIYAMA, T. & ISHIBASHI, M. 1997. Binding of high-risk human papillomavirus E6 oncoproteins to the human homologue of the Drosophila discs large tumor suppressor protein. *Proc Natl Acad Sci U S A,* 94**,** 11612-6.

KLEIN, N. P., HANSEN, J., CHAO, C., VELICER, C., EMERY, M., SLEZAK, J., LEWIS, N., DEOSARANSINGH, K., SY, L., ACKERSON, B., CHEETHAM, T. C., LIAW, K. L., TAKHAR, H. & JACOBSEN, S. J. 2012. Safety of quadrivalent human papillomavirus vaccine administered routinely to females. *Arch Pediatr Adolesc Med,* 166**,** 1140-8.

KLENOVA, E. & OHLSSON, R. 2005. Poly(ADP-ribosyl)ation and epigenetics. Is CTCF PARt of the plot? *Cell Cycle,* 4**,** 96-101.

KLENOVA, E. M., CHERNUKHIN, I. V., EL-KADY, A., LEE, R. E., PUGACHEVA, E. M., LOUKINOV, D. I., GOODWIN, G. H., DELGADO, D., FILIPPOVA, G. N., LEON, J., MORSE, H. C., 3RD, NEIMAN, P. E. & LOBANENKOV, V. V. 2001. Functional phosphorylation sites in the C-terminal region of the multivalent multifunctional transcriptional factor CTCF. *Mol Cell Biol,* 21**,** 2221-34.

KLENOVA, E. M., NICOLAS, R. H., PATERSON, H. F., CARNE, A. F., HEATH, C. M., GOODWIN, G. H., NEIMAN, P. E. & LOBANENKOV, V. V. 1993. CTCF, a conserved nuclear factor required for optimal transcriptional activity of the chicken c-myc gene, is an 11-Zn-finger protein differentially expressed in multiple forms. *Mol Cell Biol,* 13**,** 7612-24.

KLENOVA, E. M., NICOLAS, R. H., U, S., CARNE, A. F., LEE, R. E., LOBANENKOV, V. V. & GOODWIN, G. H. 1997. Molecular weight abnormalities of the CTCF transcription factor: CTCF migrates aberrantly in SDS-PAGE and the size of the expressed protein is affected by the UTRs and sequences within the coding region of the CTCF gene. *Nucleic Acids Res,* 25**,** 466-74.

KLINGELHUTZ, A. J. & ROMAN, A. 2012. Cellular transformation by human papillomaviruses: lessons learned by comparing high- and low-risk viruses. *Virology,* 424**,** 77-98.

KNIGHT, G. L., GRAINGER, J. R., GALLIMORE, P. H. & ROBERTS, S. 2004. Cooperation between different forms of the human papillomavirus type 1 E4 protein to block cell cycle progression and cellular DNA synthesis. *J Virol,* 78**,** 13920-33.

KNIGHT, G. L., PUGH, A. G., YATES, E., BELL, I., WILSON, R., MOODY, C. A., LAIMINS, L. A. & ROBERTS, S. 2011. A cyclin-binding motif in human papillomavirus type 18 (HPV18) E1^E4 is necessary for association with CDK-cyclin complexes and G2/M cell cycle arrest of keratinocytes, but is not required for differentiation-dependent viral genome amplification or L1 capsid protein expression. *Virology,* 412**,** 196-210.

KNIGHT, G. L., TURNELL, A. S. & ROBERTS, S. 2006. Role for Wee1 in inhibition of G2-to-M transition through the cooperation of distinct human papillomavirus type 1 E4 proteins. *J Virol,* 80**,** 7416-26.

KNIGHT, J. D., LI, R. & BOTCHAN, M. 1991. The activation domain of the bovine papillomavirus E2 protein mediates association of DNA-bound dimers to form DNA loops. *Proc Natl Acad Sci U S A,* 88**,** 3204-8.

KOMANDER, D. & RAPE, M. 2012. The ubiquitin code. *Annu Rev Biochem,* 81**,** 203-29.

KOSHIOL, J. E., SCHROEDER, J. C., JAMIESON, D. J., MARSHALL, S. W., DUERR, A., HEILIG, C. M., SHAH, K. V., KLEIN, R. S., CU-UVIN, S., SCHUMAN, P., CELENTANO, D. & SMITH, J. S. 2006. Time to clearance of human papillomavirus infection by type and human immunodeficiency virus serostatus. *Int J Cancer,* 119**,** 1623-9.

KOTHINTI, R., TABATABAI, N. M. & PETERING, D. H. 2011. Electrophoretic mobility shift assay of zinc finger proteins: competition for Zn(2+) bound to Sp1 in protocols including EDTA. *J Inorg Biochem,* 105**,** 569-76.

KOUTSKY, L. 1997. Epidemiology of genital human papillomavirus infection. *Am J Med,* 102**,** 3-8.

KOVELMAN, R., BILTER, G. K., GLEZER, E., TSOU, A. Y. & BARBOSA, M. S. 1996. Enhanced transcriptional activation by E2 proteins from the oncogenic human papillomaviruses. *J Virol,* 70**,** 7549-60.

KRAWCZYK, E., SUPRYNOWICZ, F. A., HEBERT, J. D., KAMONJOH, C. M. & SCHLEGEL, R. 2011. The human papillomavirus type 16 E5 oncoprotein translocates calpactin I to the perinuclear region. *J Virol,* 85**,** 10968-75.

KRAWCZYK, E., SUPRYNOWICZ, F. A., LIU, X., DAI, Y., HARTMANN, D. P., HANOVER, J. & SCHLEGEL, R. 2008. Koilocytosis: a cooperative interaction between the human papillomavirus E5 and E6 oncoproteins. *Am J Pathol,* 173**,** 682-8.

KUBAT, N. J., AMELIO, A. L., GIORDANI, N. V. & BLOOM, D. C. 2004a. The herpes simplex virus type 1 latency-associated transcript (LAT) enhancer/rcr is hyperacetylated during latency independently of LAT transcription. *J Virol,* 78**,** 12508-18.

KUBAT, N. J., TRAN, R. K., MCANANY, P. & BLOOM, D. C. 2004b. Specific histone tail modification and not DNA methylation is a determinant of herpes simplex virus type 1 latent gene expression. *J Virol,* 78**,** 1139-49.

KURG, R., TEKKEL, H., ABROI, A. & USTAV, M. 2006. Characterization of the functional activities of the bovine papillomavirus type 1 E2 protein single-chain heterodimers. *J Virol,* 80**,** 11218-25.

KURG, R., UUSEN, P., VOSA, L. & USTAV, M. 2010. Human papillomavirus E2 protein with single activation domain initiates HPV18 genome replication, but is not sufficient for long-term maintenance of virus genome. *Virology,* 408**,** 159-66.

KURUKUTI, S., TIWARI, V. K., TAVOOSIDANA, G., PUGACHEVA, E., MURRELL, A., ZHAO, Z., LOBANENKOV, V., REIK, W. & OHLSSON, R. 2006. CTCF binding at the H19 imprinting control region mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to Igf2. *Proc Natl Acad Sci U S A,* 103**,** 10684-9.

LACE, M. J., ANSON, J. R., KLINGELHUTZ, A. J., LEE, J. H., BOSSLER, A. D., HAUGEN, T. H. & TUREK, L. P. 2009. Human papillomavirus (HPV) type 18 induces extended growth in primary human cervical, tonsillar, or foreskin keratinocytes more effectively than other high-risk mucosal HPVs. *J Virol,* 83**,** 11784-94.

LACE, M. J., ANSON, J. R., THOMAS, G. S., TUREK, L. P. & HAUGEN, T. H. 2008. The E8--E2 gene product of human papillomavirus type 16 represses early transcription and replication but is dispensable for viral plasmid persistence in keratinocytes. *J Virol,* 82**,** 10841-53.

LACE, M. J., USHIKAI, M., YAMAKAWA, Y., ANSON, J. R., ISHIJI, T., TUREK, L. P. & HAUGEN, T. H. 2012. The truncated C-terminal E2 (E2-TR) protein of bovine papillomavirus (BPV) type-1 is a transactivator that modulates transcription in vivo and in vitro in a manner distinct from the E2-TA and E8^E2 gene products. *Virology,* 429**,** 99-111.

LAMBERT, P. F., MONK, B. C. & HOWLEY, P. M. 1990. Phenotypic analysis of bovine papillomavirus type 1 E2 repressor mutants. *J Virol,* 64**,** 950-6.

LAURSON, J., KHAN, S., CHUNG, R., CROSS, K. & RAJ, K. 2010. Epigenetic repression of E-cadherin by human papillomavirus 16 E7 protein. *Carcinogenesis,* 31**,** 918-26.

LEE, D., KIM, J. W., KIM, K., JOE, C. O., SCHREIBER, V., MENISSIER-DE MURCIA, J. & CHOE, J. 2002. Functional interaction between human papillomavirus type 18 E2 and poly(ADP-ribose) polymerase 1. *Oncogene,* 21**,** 5877-85.

LEE, S. S., GLAUNSINGER, B., MANTOVANI, F., BANKS, L. & JAVIER, R. T. 2000. Multi-PDZ domain protein MUPP1 is a cellular target for both adenovirus E4-ORF1 and high-risk papillomavirus type 18 E6 oncoproteins. *J Virol,* 74**,** 9680-93.

LEE, S. S., WEISS, R. S. & JAVIER, R. T. 1997. Binding of human virus oncoproteins to hDlg/SAP97, a mammalian homolog of the Drosophila discs large tumor suppressor protein. *Proc Natl Acad Sci U S A,* 94**,** 6670-5.

LEHTINEN, M., PAAVONEN, J., WHEELER, C. M., JAISAMRARN, U., GARLAND, S. M., CASTELLSAGUE, X., SKINNER, S. R., APTER, D., NAUD, P., SALMERON, J., CHOW, S. N., KITCHENER, H., TEIXEIRA, J. C., HEDRICK, J., LIMSON, G., SZAREWSKI, A., ROMANOWSKI, B., AOKI, F. Y., SCHWARZ, T. F., POPPE, W. A., DE CARVALHO, N. S., GERMAR, M. J., PETERS, K., MINDEL, A., DE SUTTER, P., BOSCH, F. X., DAVID, M. P., DESCAMPS, D., STRUYF, F. & DUBIN, G. 2012. Overall efficacy of HPV-16/18 AS04-adjuvanted vaccine against grade 3 or greater cervical intraepithelial neoplasia: 4-year end-of-study analysis of the randomised, double-blind PATRICIA trial. *Lancet Oncol,* 13**,** 89-99.

LEONARD, S. M., WEI, W., COLLINS, S. I., PEREIRA, M., DIYAF, A., CONSTANDINOU-WILLIAMS, C., YOUNG, L. S., ROBERTS, S. & WOODMAN, C. B. 2012. Oncogenic human papillomavirus imposes an instructive pattern of DNA methylation changes which parallel the natural history of cervical HPV infection in young women. *Carcinogenesis,* 33**,** 1286-93.

LI, H., YANG, Y., ZHANG, R., CAI, Y., YANG, X., WANG, Z., LI, Y., CHENG, X., YE, X., XIANG, Y. & ZHU, B. 2013. Preferential sites for the integration and disruption of human papillomavirus 16 in cervical lesions. *J Clin Virol,* 56**,** 342-7.

LI, R., KNIGHT, J., BREAM, G., STENLUND, A. & BOTCHAN, M. 1989. Specific recognition nucleotides and their DNA context determine the affinity of E2 protein for 17 binding sites in the BPV-1 genome. *Genes Dev,* 3**,** 510-26.

LIEBERMAN-AIDEN, E., VAN BERKUM, N. L., WILLIAMS, L., IMAKAEV, M., RAGOCZY, T., TELLING, A., AMIT, I., LAJOIE, B. R., SABO, P. J., DORSCHNER, M. O., SANDSTROM, R., BERNSTEIN, B., BENDER, M. A., GROUDINE, M., GNIRKE, A., STAMATOYANNOPOULOS, J., MIRNY, L. A., LANDER, E. S. & DEKKER, J. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science,* 326**,** 289-93.

LIM, D. A., GOSSEN, M., LEHMAN, C. W. & BOTCHAN, M. R. 1998. Competition for DNA binding sites between the short and long forms of E2 dimers underlies repression in bovine papillomavirus type 1 DNA replication control. *J Virol,* 72**,** 1931-40.

LING, J. Q., LI, T., HU, J. F., VU, T. H., CHEN, H. L., QIU, X. W., CHERRY, A. M. & HOFFMAN, A. R. 2006. CTCF mediates interchromosomal colocalization between Igf2/H19 and Wsb1/Nf1. *Science,* 312**,** 269-72.

LIVAK, K. J. & SCHMITTGEN, T. D. 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods,* 25**,** 402-8.

LOBANENKOV, V. V., NICOLAS, R. H., ADLER, V. V., PATERSON, H., KLENOVA, E. M., POLOTSKAJA, A. V. & GOODWIN, G. H. 1990. A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5'-flanking sequence of the chicken c-myc gene. *Oncogene,* 5**,** 1743-53.

LOUKINOV, D. I., PUGACHEVA, E., VATOLIN, S., PACK, S. D., MOON, H., CHERNUKHIN, I., MANNAN, P., LARSSON, E., KANDURI, C., VOSTROV, A. A., CUI, H., NIEMITZ, E. L., RASKO, J. E., DOCQUIER, F. M., KISTLER, M., BREEN, J. J., ZHUANG, Z., QUITSCHKE, W. W., RENKAWITZ, R., KLENOVA, E. M., FEINBERG, A. P., OHLSSON, R., MORSE, H. C., 3RD & LOBANENKOV, V. V. 2002. BORIS, a novel male germ-line-specific protein associated with epigenetic reprogramming events, shares the same 11-zinc-finger

domain with CTCF, the insulator protein involved in reading imprinting marks in the soma. *Proc Natl Acad Sci U S A,* 99**,** 6806-11.

LOWY, D. R. & SCHILLER, J. T. 2006. Prophylactic human papillomavirus vaccines. *J Clin Invest,* 116**,** 1167-73.

MACHALEK, D. A., POYNTEN, M., JIN, F., FAIRLEY, C. K., FARNSWORTH, A., GARLAND, S. M., HILLMAN, R. J., PETOUMENOS, K., ROBERTS, J., TABRIZI, S. N., TEMPLETON, D. J. & GRULICH, A. E. 2012. Anal human papillomavirus infection and associated neoplastic lesions in men who have sex with men: a systematic review and meta-analysis. *Lancet Oncol,* 13**,** 487-500.

MACPHERSON, M. J., BEATTY, L. G., ZHOU, W., DU, M. & SADOWSKI, P. D. 2009. The CTCF insulator protein is posttranslationally modified by SUMO. *Mol Cell Biol,* 29**,** 714-25.

MACPHERSON, M. J. & SADOWSKI, P. D. 2010. The CTCF insulator protein forms an unusual DNA structure. *BMC Mol Biol,* 11**,** 101.

MAGLENNON, G. A., MCINTOSH, P. & DOORBAR, J. 2011. Persistence of viral DNA in the epithelial basal layer suggests a model for papillomavirus latency following immune regression. *Virology,* 414**,** 153-63.

MARKOWITZ, L. 2011. Recommendations on the use of quadrivalent human papillomavirus vaccine in males--Advisory Committee on Immunization Practices (ACIP), 2011. *MMWR Morb Mortal Wkly Rep,* 60**,** 1705-8.

MASTERSON, P. J., STANLEY, M. A., LEWIS, A. P. & ROMANOS, M. A. 1998. A C-terminal helicase domain of the human papillomavirus E1 protein binds E2 and the DNA polymerase alpha-primase p68 subunit. *J Virol,* 72**,** 7407-19.

MATSUKURA, T., KOI, S. & SUGASE, M. 1989. Both episomal and integrated forms of human papillomavirus type 16 are involved in invasive cervical cancers. *Virology,* 172**,** 63-72.

MAUFORT, J. P., SHAI, A., PITOT, H. C. & LAMBERT, P. F. 2010. A role for HPV16 E5 in cervical carcinogenesis. *Cancer Res,* 70**,** 2924-31.

MAUNAKEA, A. K., CHEPELEV, I., CUI, K. & ZHAO, K. 2013. Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition. *Cell Res*, 11, 1256-69.

MCBRIDE, A. A. 2008. Replication and partitioning of papillomavirus genomes. *Adv Virus Res,* 72**,** 155-205.

MCBRIDE, A. A. 2013. The Papillomavirus E2 proteins. *Virology*, 445, 57-79.

MCBRIDE, A. A., BYRNE, J. C. & HOWLEY, P. M. 1989. E2 polypeptides encoded by bovine papillomavirus type 1 form dimers through the common carboxyl-terminal domain: transactivation is mediated by the conserved amino-terminal domain. *Proc Natl Acad Sci U S A,* 86**,** 510-4.

MCBRIDE, A. A., OLIVEIRA, J. G. & MCPHILLIPS, M. G. 2006. Partitioning viral genomes in mitosis: same idea, different targets. *Cell Cycle,* 5**,** 1499-502.

MCINTOSH, P. B., MARTIN, S. R., JACKSON, D. J., KHAN, J., ISAACSON, E. R., CALDER, L., RAJ, K., GRIFFIN, H. M., WANG, Q., LASKEY, P., ECCLESTON, J. F. & DOORBAR, J. 2008. Structural analysis reveals an amyloid form of the human papillomavirus type 16 E1--E4 protein and provides a molecular basis for its accumulation. *J Virol,* 82**,** 8196-203.

MCLAUGHLIN-DRUBIN, M. E., CRUM, C. P. & MUNGER, K. 2011. Human papillomavirus E7 oncoprotein induces KDM6A and KDM6B histone demethylase expression and causes epigenetic reprogramming. *Proc Natl Acad Sci U S A,* 108**,** 2130-5.

MCMURRAY, H. R. & MCCANCE, D. J. 2003. Human papillomavirus type 16 E6 activates TERT gene transcription through induction of c-Myc and release of USF-mediated repression. *J Virol,* 77**,** 9852-61.

MCPHILLIPS, M. G., OLIVEIRA, J. G., SPINDLER, J. E., MITRA, R. & MCBRIDE, A. A. 2006. Brd4 is required for e2-mediated transcriptional activation but not genome partitioning of all papillomaviruses. *J Virol,* 80**,** 9530-43.

MELSHEIMER, P., VINOKUROVA, S., WENTZENSEN, N., BASTERT, G. & VON KNEBEL DOEBERITZ, M. 2004. DNA aneuploidy and integration of human papillomavirus type 16 e6/e7 oncogenes in intraepithelial neoplasia and invasive squamous cell carcinoma of the cervix uteri. *Clin Cancer Res,* 10**,** 3059-63.

METZGER, M. B., HRISTOVA, V. A. & WEISSMAN, A. M. 2012. HECT and RING finger families of E3 ubiquitin ligases at a glance. *J Cell Sci,* 125**,** 531-7.

MIDDLETON, K., PEH, W., SOUTHERN, S., GRIFFIN, H., SOTLAR, K., NAKAHARA, T., EL-SHERIF, A., MORRIS, L., SETH, R., HIBMA, M., JENKINS, D., LAMBERT, P., COLEMAN, N. & DOORBAR, J. 2003. Organization of human papillomavirus productive cycle during neoplastic progression provides a basis for selection of diagnostic markers. *J Virol,* 77**,** 10186-201.

MILLS, S. A. & MARLETTA, M. A. 2005. Metal binding characteristics and role of iron oxidation in the ferric uptake regulator from Escherichia coli. *Biochemistry,* 44**,** 13553-9.

MOHR, I. J., CLARK, R., SUN, S., ANDROPHY, E. J., MACPHERSON, P. & BOTCHAN, M. R. 1990. Targeting the E1 replication protein to the papillomavirus origin of replication by complex formation with the E2 transactivator. *Science,* 250**,** 1694-9.

MOODLEY, J. 2004. Combined oral contraceptives and cervical cancer. *Curr Opin Obstet Gynecol,* 16**,** 27-9.

MUKHOPADHYAY, D. & RIEZMAN, H. 2007. Proteasome-independent functions of ubiquitin in endocytosis and signaling. *Science,* 315**,** 201-5.

MUNOZ, N., BOSCH, F. X., CASTELLSAGUE, X., DIAZ, M., DE SANJOSE, S., HAMMOUDA, D., SHAH, K. V. & MEIJER, C. J. 2004. Against which human papillomavirus types shall we vaccinate and screen? The international perspective. *Int J Cancer,* 111**,** 278-85.

MUNOZ, N., BOSCH, F. X., DE SANJOSE, S., HERRERO, R., CASTELLSAGUE, X., SHAH, K. V., SNIJDERS, P. J. & MEIJER, C. J. 2003. Epidemiologic classification of human papillomavirus types associated with cervical cancer. *N Engl J Med,* 348**,** 518-27.

MUNOZ, N., CASTELLSAGUE, X., DE GONZALEZ, A. B. & GISSMANN, L. 2006. Chapter 1: HPV in the etiology of human cancer. *Vaccine,* 24 Suppl 3**,** S3/1-10.

NAGARAJAN, P., CHIN, S. S., WANG, D., LIU, S., SINHA, S. & GARRETT-SINHA, L. A. 2010. Ets1 blocks terminal differentiation of keratinocytes and induces expression of matrix metalloproteases and innate immune mediators. *J Cell Sci,* 123**,** 3566-75.

NAIR, P., SOMASUNDARAM, K. & KRISHNA, S. 2003. Activated Notch1 inhibits p53-induced apoptosis and sustains transformation by human papillomavirus type 16 E6 and E7 oncogenes through a PI3K-PKB/Akt-dependent pathway. *J Virol,* 77**,** 7106-12.

NAKAGAWA, S. & HUIBREGTSE, J. M. 2000. Human scribble (Vartul) is targeted for ubiquitin-mediated degradation by the high-risk papillomavirus E6 proteins and the E6AP ubiquitin-protein ligase. *Mol Cell Biol,* 20**,** 8244-53.

NAKAHARA, T., PEH, W. L., DOORBAR, J., LEE, D. & LAMBERT, P. F. 2005. Human papillomavirus type 16 E1circumflexE4 contributes to multiple facets of the papillomavirus life cycle. *J Virol,* 79**,** 13150-65.

NAKAHASHI, H., KWON, K. R., RESCH, W., VIAN, L., DOSE, M., STAVREVA, D., HAKIM, O., PRUETT, N., NELSON, S., YAMANE, A., QIAN, J., DUBOIS, W., WELSH, S., PHAIR, R. D., PUGH, B. F., LOBANENKOV, V., HAGER, G. L. & CASELLAS, R. 2013. A Genome-wide Map of CTCF Multivalency Redefines the CTCF Code. *Cell Rep,* 3**,** 1678-89.

NANBO, A., SUGDEN, A. & SUGDEN, B. 2007. The coupling of synthesis and partitioning of EBV's plasmid replicon is revealed in live cells. *EMBO J,* 26**,** 4252-62.

NARAHARI, J., FISK, J. C., MELENDY, T. & ROMAN, A. 2006. Interactions of the cellular CCAAT displacement protein and human papillomavirus E2 protein with the viral origin of replication can regulate DNA replication. *Virology,* 350**,** 302-11.

NASMYTH, K. & HAERING, C. H. 2005. The structure and function of SMC and kleisin complexes. *Annu Rev Biochem,* 74**,** 595-648.

NATIVIO, R., WENDT, K. S., ITO, Y., HUDDLESTON, J. E., URIBE-LEWIS, S., WOODFINE, K., KRUEGER, C., REIK, W., PETERS, J. M. & MURRELL, A. 2009. Cohesin is required for higher-order chromatin conformation at the imprinted IGF2-H19 locus. *PLoS Genet,* 5**,** e1000739.

NEES, M., GEOGHEGAN, J. M., HYMAN, T., FRANK, S., MILLER, L. & WOODWORTH, C. D. 2001. Papillomavirus type 16 oncogenes downregulate expression of interferon-responsive genes and upregulate proliferation-associated and NF-kappaB-responsive genes in cervical keratinocytes. *J Virol,* 75**,** 4283-96.

NGUYEN, D. X., BAGLIA, L. A., HUANG, S. M., BAKER, C. M. & MCCANCE, D. J. 2004. Acetylation regulates the differentiation-specific functions of the retinoblastoma protein. *EMBO J,* 23**,** 1609-18.

NGUYEN, D. X. & MCCANCE, D. J. 2005. Role of the retinoblastoma tumor suppressor protein in cellular differentiation. *J Cell Biochem,* 94**,** 870-9.

NICHOLLS, P. K., KLAUNBERG, B. A., MOORE, R. A., SANTOS, E. B., PARRY, N. R., GOUGH, G. W. & STANLEY, M. A. 1999. Naturally occurring, nonregressing canine oral papillomavirus infection: host immunity, virus characterization, and experimental infection. *Virology,* 265**,** 365-74.

NICOLAS, M., WOLFER, A., RAJ, K., KUMMER, J. A., MILL, P., VAN NOORT, M., HUI, C. C., CLEVERS, H., DOTTO, G. P. & RADTKE, F. 2003. Notch1 functions as a tumor suppressor in mouse skin. *Nat Genet,* 33**,** 416-21.

NINDL, I., GOTTSCHLING, M. & STOCKFLETH, E. 2007. Human papillomaviruses and non-melanoma skin cancer: basic virology and clinical manifestations. *Dis Markers,* 23**,** 247-59.

NOYA, F., CHIEN, W. M., BROKER, T. R. & CHOW, L. T. 2001. p21cip1 Degradation in differentiated keratinocytes is abrogated by costabilization with cyclin E induced by human papillomavirus E7. *J Virol,* 75**,** 6121-34.

OHLSSON, R., BARTKUHN, M. & RENKAWITZ, R. 2010. CTCF shapes chromatin by multiple mechanisms: the impact of 20 years of CTCF research on understanding the workings of chromatin. *Chromosoma,* 119**,** 351-60.

OLIVEIRA, J. G., COLF, L. A. & MCBRIDE, A. A. 2006. Variations in the association of papillomavirus E2 proteins with mitotic chromosomes. *Proc Natl Acad Sci U S A,* 103**,** 1047-52.

OZSARAN, A. A., ATES, T., DIKMEN, Y., ZEYTINOGLU, A., TEREK, C., ERHAN, Y., OZACAR, T. & BILGIC, A. 1999. Evaluation of the risk of cervical intraepithelial neoplasia and human papilloma virus infection in renal transplant patients receiving immunosuppressive therapy. *Eur J Gynaecol Oncol,* 20**,** 127-30.

PAAVONEN, J., JENKINS, D., BOSCH, F. X., NAUD, P., SALMERON, J., WHEELER, C. M., CHOW, S. N., APTER, D. L., KITCHENER, H. C., CASTELLSAGUE, X., DE CARVALHO, N. S., SKINNER, S. R., HARPER, D. M., HEDRICK, J. A., JAISAMRARN, U., LIMSON, G. A., DIONNE, M., QUINT, W., SPIESSENS, B., PEETERS, P., STRUYF, F., WIETING, S. L., LEHTINEN, M. O. & DUBIN, G. 2007. Efficacy of a prophylactic adjuvanted bivalent L1 virus-like-particle vaccine against infection with human papillomavirus types 16 and 18 in young women: an interim analysis of a phase III double-blind, randomised controlled trial. *Lancet,* 369**,** 2161-70.

PAAVONEN, J., NAUD, P., SALMERON, J., WHEELER, C. M., CHOW, S. N., APTER, D., KITCHENER, H., CASTELLSAGUE, X., TEIXEIRA, J. C., SKINNER, S. R., HEDRICK, J., JAISAMRARN, U., LIMSON, G., GARLAND, S., SZAREWSKI, A., ROMANOWSKI, B., AOKI, F. Y., SCHWARZ, T. F., POPPE, W. A., BOSCH, F. X., JENKINS, D., HARDT, K., ZAHAF, T., DESCAMPS, D., STRUYF, F., LEHTINEN, M. & DUBIN, G. 2009. Efficacy of human papillomavirus (HPV)-16/18 AS04-adjuvanted vaccine against cervical infection and precancer caused by oncogenic HPV types (PATRICIA): final analysis of a double-blind, randomised study in young women. *Lancet,* 374**,** 301-14.

PALSTRA, R. J., TOLHUIS, B., SPLINTER, E., NIJMEIJER, R., GROSVELD, F. & DE LAAT, W. 2003. The beta-globin nuclear compartment in development and erythroid differentiation. *Nat Genet,* 35**,** 190-4.

PAPANICOLAOU, G. N. 1948. The cell smear method of diagnosing cancer. *Am J Public Health Nations Health,* 38**,** 202-5.

PARELHO, V., HADJUR, S., SPIVAKOV, M., LELEU, M., SAUER, S., GREGSON, H. C., JARMUZ, A., CANZONETTA, C., WEBSTER, Z., NESTEROVA, T., COBB, B. S., YOKOMORI, K., DILLON, N., ARAGON, L., FISHER, A. G. & MERKENSCHLAGER, M. 2008. Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell,* 132**,** 422-33.

PARISH, J. L., BEAN, A. M., PARK, R. B. & ANDROPHY, E. J. 2006. ChlR1 is required for loading papillomavirus E2 onto mitotic chromosomes and viral genome maintenance. *Mol Cell,* 24**,** 867-76.

PARK, R. B. & ANDROPHY, E. J. 2002. Genetic analysis of high-risk e6 in episomal maintenance of human papillomavirus genomes in primary human keratinocytes. *J Virol,* 76**,** 11359-64.

PARKIN, D. M. 2011. 11. Cancers attributable to infection in the UK in 2010. *Br J Cancer,* 105 Suppl 2**,** S49-56.

PARKIN, D. M., BRAY, F., FERLAY, J. & PISANI, P. 2005. Global cancer statistics, 2002. *CA Cancer J Clin,* 55**,** 74-108.

PATERNOSTER, D. M., CESTER, M., RESENTE, C., PASCOLI, I., NANHORNGUE, K., MARCHINI, F., BOCCAGNI, P., CILLO, U., RIBALDONE, R., AMORUSO, E., COCCA, N., CUCCOLO, V., BERTOLINO, M., SURICO, N. & STRATTA, P. 2008. Human papilloma virus infection and cervical intraepithelial neoplasia in transplanted patients. *Transplant Proc,* 40**,** 1877-80.

PATTURAJAN, M., NOMOTO, S., SOMMER, M., FOMENKOV, A., HIBI, K., ZANGEN, R., POLIAK, N., CALIFANO, J., TRINK, B., RATOVITSKI, E. & SIDRANSKY, D. 2002. DeltaNp63 induces beta-catenin nuclear accumulation and signaling. *Cancer Cell,* 1**,** 369-79.

PEH, W. L., BRANDSMA, J. L., CHRISTENSEN, N. D., CLADEL, N. M., WU, X. & DOORBAR, J. 2004. The viral E4 protein is required for the completion of the cottontail rabbit papillomavirus productive cycle in vivo. *J Virol,* 78**,** 2142-51.

PENROSE, K. J. & MCBRIDE, A. A. 2000. Proteasome-mediated degradation of the papillomavirus E2-TA protein is regulated by phosphorylation and can modulate viral genome copy number. *J Virol,* 74**,** 6031-8.

PEREA, S. E., MASSIMI, P. & BANKS, L. 2000. Human papillomavirus type 16 E7 impairs the activation of the interferon regulatory factor-1. *Int J Mol Med,* 5**,** 661-6.

PERIC-HUPKES, D., MEULEMAN, W., PAGIE, L., BRUGGEMAN, S. W., SOLOVEI, I., BRUGMAN, W., GRAF, S., FLICEK, P., KERKHOVEN, R. M., VAN LOHUIZEN, M., REINDERS, M., WESSELS, L. & VAN STEENSEL, B. 2010. Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Mol Cell,* 38**,** 603-13.

PETT, M. & COLEMAN, N. 2007. Integration of high-risk human papillomavirus: a key event in cervical carcinogenesis? *J Pathol,* 212**,** 356-67.

PFAFFL, M. W. 2001. A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res,* 29**,** e45.

PHILLIPS, J. E. & CORCES, V. G. 2009. CTCF: master weaver of the genome. *Cell,* 137**,** 1194-211.

PICKART, C. M. & EDDINS, M. J. 2004. Ubiquitin: structures, functions, mechanisms. *Biochim Biophys Acta,* 1695**,** 55-72.

PIIRSOO, M., USTAV, E., MANDEL, T., STENLUND, A. & USTAV, M. 1996. Cis and trans requirements for stable episomal maintenance of the BPV-1 replicator. *EMBO J,* 15**,** 1-11.

PIM, D. & BANKS, L. 2010. Interaction of viral oncoproteins with cellular target molecules: infection with high-risk vs low-risk human papillomaviruses. *APMIS,* 118**,** 471-93.

PISANI, P., BRAY, F. & PARKIN, D. M. 2002. Estimates of the world-wide prevalence of cancer for 25 sites in the adult population. *Int J Cancer,* 97**,** 72-81.

PROBST, A. V., DUNLEAVY, E. & ALMOUZNI, G. 2009. Epigenetic inheritance during the cell cycle. *Nat Rev Mol Cell Biol,* 10**,** 192-206.

PROWELLER, A., TU, L., LEPORE, J. J., CHENG, L., LU, M. M., SEYKORA, J., MILLAR, S. E., PEAR, W. S. & PARMACEK, M. S. 2006. Impaired notch signaling promotes de novo squamous cell carcinoma formation. *Cancer Res,* 66**,** 7438-44.

PYEON, D., PEARCE, S. M., LANK, S. M., AHLQUIST, P. & LAMBERT, P. F. 2009. Establishment of human papillomavirus infection requires cell cycle progression. *PLoS Pathog,* 5**,** e1000318.

QUINT, W., JENKINS, D., MOLIJN, A., STRUIJK, L., VAN DE SANDT, M., DOORBAR, J., MOLS, J., VAN HOOF, C., HARDT, K., STRUYF, F. & COLAU, B. 2012. One virus, one lesion--individual components of CIN lesions contain a specific HPV type. *J Pathol,* 227**,** 62-71.

RANGARAJAN, A., SYAL, R., SELVARAJAH, S., CHAKRABARTI, O., SARIN, A. & KRISHNA, S. 2001. Activated Notch1 signaling cooperates with papillomavirus oncogenes in transformation and generates resistance to apoptosis on matrix withdrawal through PKB/Akt. *Virology,* 286**,** 23-30.

RECILLAS-TARGA, F., PIKAART, M. J., BURGESS-BEUSSE, B., BELL, A. C., LITT, M. D., WEST, A. G., GASZNER, M. & FELSENFELD, G. 2002. Position-effect protection and enhancer blocking by the chicken beta-globin insulator are separable activities. *Proc Natl Acad Sci U S A,* 99**,** 6883-8.

RICHARDSON, H., KELSALL, G., TELLIER, P., VOYER, H., ABRAHAMOWICZ, M., FERENCZY, A., COUTLEE, F. & FRANCO, E. L. 2003. The natural history of type-specific human papillomavirus infections in female university students. *Cancer Epidemiol Biomarkers Prev,* 12**,** 485-90.

RIESE, D. J., 2ND, SETTLEMAN, J., NEARY, K. & DIMAIO, D. 1990. Bovine papillomavirus E2 repressor mutant displays a high-copy-number phenotype and enhanced transforming activity. *J Virol,* 64**,** 944-9.

ROMAN, A. 2006. The human papillomavirus E7 protein shines a spotlight on the pRB family member, p130. *Cell Cycle,* 5**,** 567-8.

ROSA-GARRIDO, M., CEBALLOS, L., ALONSO-LECUE, P., ABRAIRA, C., DELGADO, M. D. & GANDARILLAS, A. 2012. A cell cycle role for the epigenetic factor CTCF-L/BORIS. *PLoS One,* 7**,** e39371.

ROSS-INNES, C. S., BROWN, G. D. & CARROLL, J. S. 2011. A co-ordinated interaction between CTCF and ER in breast cancer cells. *BMC Genomics,* 12**,** 593.

RUBIO, E. D., REISS, D. J., WELCSH, P. L., DISTECHE, C. M., FILIPPOVA, G. N., BALIGA, N. S., AEBERSOLD, R., RANISH, J. A. & KRUMM, A. 2008. CTCF physically links cohesin to chromatin. *Proc Natl Acad Sci U S A,* 105**,** 8309-14.

RUSSELL, M., RAHEJA, V. & JAIYESIMI, R. 2013. Human papillomavirus vaccination in adolescence. *Perspect Public Health*, 133, 320-4.

SALAMON, D., BANATI, F., KOROKNAI, A., RAVASZ, M., SZENTHE, K., BATHORI, Z., BAKOS, A., NILLER, H. H., WOLF, H. & MINAROVITS, J. 2009. Binding of CCCTC-binding factor in vivo to the region located between Rep* and the C promoter of Epstein-Barr virus is unaffected by CpG methylation and does not correlate with Cp activity. *J Gen Virol,* 90**,** 1183-9.

SAMBROOK, J. & RUSSELL, D. W. 2006. Gel Retardation Assays for DNA-binding Proteins. *CSH Protoc,* 2006.

SASLOW, D., SOLOMON, D., LAWSON, H. W., KILLACKEY, M., KULASINGAM, S. L., CAIN, J. M., GARCIA, F. A., MORIARTY, A. T., WAXMAN, A. G., WILBUR, D. C., WENTZENSEN, N., DOWNS, L. S., JR., SPITZER, M., MOSCICKI, A. B., FRANCO, E. L., STOLER, M. H., SCHIFFMAN, M., CASTLE, P. E., MYERS, E. R., CHELMOW, D., HERZIG, A., KIM, J. J., KINNEY, W., HERSCHEL, W. L. & WALDMAN, J. 2012. American Cancer Society, American Society for Colposcopy and Cervical Pathology, and American Society for Clinical Pathology screening guidelines for the prevention and early detection of cervical cancer. *J Low Genit Tract Dis,* 16**,** 175-204.

SCHIFFMAN, M., CASTLE, P. E., JERONIMO, J., RODRIGUEZ, A. C. & WACHOLDER, S. 2007. Human papillomavirus and cervical cancer. *Lancet,* 370**,** 890-907.

SCHIFFMAN, M., CLIFFORD, G. & BUONAGURO, F. M. 2009. Classification of weakly carcinogenic human papillomavirus types: addressing the limits of epidemiology at the borderline. *Infect Agent Cancer,* 4**,** 8.

SCHIFFMAN, M., RODRIGUEZ, A. C., CHEN, Z., WACHOLDER, S., HERRERO, R., HILDESHEIM, A., DESALLE, R., BEFANO, B., YU, K., SAFAEIAN, M., SHERMAN, M. E., MORALES, J., GUILLEN, D., ALFARO, M., HUTCHINSON, M., SOLOMON, D., CASTLE, P. E. & BURK, R. D. 2010. A population-based prospective study of carcinogenic human papillomavirus variant lineages, viral persistence, and cervical neoplasia. *Cancer Res,* 70**,** 3159-69.

SCHILLER, J. T., CASTELLSAGUE, X. & GARLAND, S. M. 2012. A review of clinical trials of human papillomavirus prophylactic vaccines. *Vaccine,* 30 Suppl 5**,** F123-38.

SCHILLER, J. T., DAY, P. M. & KINES, R. C. 2010. Current understanding of the mechanism of HPV infection. *Gynecol Oncol,* 118**,** S12-7.

SCHMIDT, D., SCHWALIE, P. C., WILSON, M. D., BALLESTER, B., GONCALVES, A., KUTTER, C., BROWN, G. D., MARSHALL, A., FLICEK, P. & ODOM, D. T. 2012. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell,* 148**,** 335-48.

SCHMIEDESKAMP, M. R. & KOCKLER, D. R. 2006. Human papillomavirus vaccines. *Ann Pharmacother,* 40**,** 1344-52.

SCHUELKE, M. 2000. An economic method for the fluorescent labeling of PCR fragments. *Nat Biotechnol,* 18**,** 233-4.

SCHULMAN, B. A. & HARPER, J. W. 2009. Ubiquitin-like protein activation by E1 enzymes: the apex for downstream signalling pathways. *Nat Rev Mol Cell Biol,* 10**,** 319-31.

SCHWARZ, E., FREESE, U. K., GISSMANN, L., MAYER, W., ROGGENBUCK, B., STREMLAU, A. & ZUR HAUSEN, H. 1985. Structure and transcription of human papillomavirus sequences in cervical carcinoma cells. *Nature,* 314**,** 111-4.

SEDMAN, J. & STENLUND, A. 1995. Co-operative interaction between the initiator E1 and the transcriptional activator E2 is required for replicator specific DNA replication of bovine papillomavirus in vivo and in vitro. *EMBO J,* 14**,** 6218-28.

SEITAN, V. C., HAO, B., TACHIBANA-KONWALSKI, K., LAVAGNOLLI, T., MIRA-BONTENBAL, H., BROWN, K. E., TENG, G., CARROLL, T., TERRY, A., HORAN, K., MARKS, H., ADAMS, D. J., SCHATZ, D. G., ARAGON, L., FISHER, A. G., KRANGEL, M. S., NASMYTH, K. & MERKENSCHLAGER, M. 2011. A role for cohesin in T-cell-receptor rearrangement and thymocyte differentiation. *Nature,* 476**,** 467-71.

SENOO, M., PINTO, F., CRUM, C. P. & MCKEON, F. 2007. p63 Is essential for the proliferative potential of stem cells in stratified epithelia. *Cell,* 129**,** 523-36.

SHEN, Y., YUE, F., MCCLEARY, D. F., YE, Z., EDSALL, L., KUAN, S., WAGNER, U., DIXON, J., LEE, L., LOBANENKOV, V. V. & REN, B. 2012. A map of the cis-regulatory sequences in the mouse genome. *Nature,* 488**,** 116-20.

SHERMAN, L., ITZHAKI, H., JACKMAN, A., CHEN, J. J., KOVAL, D. & SCHLEGEL, R. 2002. Inhibition of serum- and calcium-induced terminal differentiation of human keratinocytes by HPV 16 E6: study of the association with p53 degradation, inhibition of p53 transactivation, and binding to E6BP. *Virology,* 292**,** 309-20.

SHIH, S. C., SLOPER-MOULD, K. E. & HICKE, L. 2000. Monoubiquitin carries a novel internalization signal that is appended to activated receptors. *EMBO J,* 19**,** 187-98.

SHUKLA, S., KAVAK, E., GREGORY, M., IMASHIMIZU, M., SHUTINOSKI, B., KASHLEV, M., OBERDOERFFER, P., SANDBERG, R. & OBERDOERFFER, S. 2011. CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature,* 479**,** 74-9.

SICHERO, L., SOBRINHO, J. S. & VILLA, L. L. 2012. Identification of novel cellular transcription factors that regulate early promoters of human papillomavirus types 18 and 16. *J Infect Dis,* 206**,** 867-74.

SIDDIQUI, M. A. & PERRY, C. M. 2006. Human papillomavirus quadrivalent (types 6, 11, 16, 18) recombinant vaccine (Gardasil). *Drugs,* 66**,** 1263-71; discussion 1272-3.

SIM, J., OZGUR, S., LIN, B. Y., YU, J. H., BROKER, T. R., CHOW, L. T. & GRIFFITH, J. 2008. Remodeling of the human papillomavirus type 11 replication origin into discrete nucleoprotein particles and looped structures by the E2 protein. *J Mol Biol,* 375**,** 1165-77.

SINGH, P., LEE, D. H. & SZABO, P. E. 2012. More than insulator: multiple roles of CTCF at the H19-Igf2 imprinted domain. *Front Genet,* 3**,** 214.

SMAL, C., WETZLER, D. E., DANTUR, K. I., CHEMES, L. B., GARCIA-ALAI, M. M., DELLAROLE, M., ALONSO, L. G., GASTON, K. & DE PRAT-GAY, G. 2009. The human papillomavirus E7-E2 interaction mechanism in vitro reveals a finely tuned system for modulating available E7 and E2 proteins. *Biochemistry,* 48**,** 11939-49.

STACEY, S. N., JORDAN, D., WILLIAMSON, A. J., BROWN, M., COOTE, J. H. & ARRAND, J. R. 2000. Leaky scanning is the predominant mechanism for translation of human papillomavirus type 16 E7 oncoprotein from E6/E7 bicistronic mRNA. *J Virol,* 74**,** 7284-97.

STANLEY, M. 2006. Immune responses to human papillomavirus. *Vaccine,* 24 Suppl 1**,** S16-22.

STANLEY, M. A. 2009. Immune responses to human papilloma viruses. *Indian J Med Res,* 130**,** 266-76.

STANLEY, M. A. 2012. Epithelial cell responses to infection with human papillomavirus. *Clin Microbiol Rev,* 25**,** 215-22.

STEDMAN, W., KANG, H., LIN, S., KISSIL, J. L., BARTOLOMEI, M. S. & LIEBERMAN, P. M. 2008. Cohesins localize with CTCF at the KSHV latency control region and at cellular c-myc and H19/Igf2 insulators. *EMBO J,* 27**,** 654-66.

STEGER, G. & CORBACH, S. 1997. Dose-dependent regulation of the early promoter of human papillomavirus type 18 by the viral E2 protein. *J Virol,* 71**,** 50-8.

STOPPLER, M. C., STRAIGHT, S. W., TSAO, G., SCHLEGEL, R. & MCCANCE, D. J. 1996. The E5 gene of HPV-16 enhances keratinocyte immortalization by full-length DNA. *Virology,* 223**,** 251-4.

STRAIGHT, S. W., HINKLE, P. M., JEWERS, R. J. & MCCANCE, D. J. 1993. The E5 oncoprotein of human papillomavirus type 16 transforms fibroblasts and effects the downregulation of the epidermal growth factor receptor in keratinocytes. *J Virol,* 67**,** 4521-32.

STUBENRAUCH, F., HUMMEL, M., IFTNER, T. & LAIMINS, L. A. 2000. The E8E2C protein, a negative regulator of viral transcription and replication, is required for extrachromosomal maintenance of human papillomavirus type 31 in keratinocytes. *J Virol,* 74**,** 1178-86.

STUBENRAUCH, F., LIM, H. B. & LAIMINS, L. A. 1998. Differential requirements for conserved E2 binding sites in the life cycle of oncogenic human papillomavirus type 31. *J Virol,* 72**,** 1071-7.

SZABO, P. E., TANG, S. H., SILVA, F. J., TSARK, W. M. & MANN, J. R. 2004. Role of CTCF binding sites in the Igf2/H19 imprinting control region. *Mol Cell Biol,* 24**,** 4791-800.

SZAREWSKI, A., POPPE, W. A., SKINNER, S. R., WHEELER, C. M., PAAVONEN, J., NAUD, P., SALMERON, J., CHOW, S. N., APTER, D., KITCHENER, H., CASTELLSAGUE, X., TEIXEIRA, J. C., HEDRICK, J., JAISAMRARN, U., LIMSON, G., GARLAND, S., ROMANOWSKI, B., AOKI, F. Y., SCHWARZ, T. F., BOSCH, F. X., HARPER, D. M., HARDT, K., ZAHAF, T., DESCAMPS, D., STRUYF, F., LEHTINEN, M. & DUBIN, G. 2012. Efficacy of the human papillomavirus (HPV)-16/18 AS04-adjuvanted vaccine in women aged 15-25 years with and without serological evidence of previous exposure to HPV-16/18. *Int J Cancer,* 131**,** 106-16.

TALORA, C., SGROI, D. C., CRUM, C. P. & DOTTO, G. P. 2002. Specific down-modulation of Notch1 signaling in cervical cancer cells is required for sustained HPV-E6/E7 expression and late steps of malignant transformation. *Genes Dev,* 16**,** 2252-63.

TAN, C. L., GUNARATNE, J., LAI, D., CARTHAGENA, L., WANG, Q., XUE, Y. Z., QUEK, L. S., DOORBAR, J., BACHELERIE, F., THIERRY, F. & BELLANGER, S. 2012. HPV-18 E2^E4 chimera: 2 new spliced transcripts and proteins induced by keratinocyte differentiation. *Virology,* 429**,** 47-56.

TANAY, A. 2006. Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Res,* 16**,** 962-72.

TEMPERA, I., WIEDMER, A., DHEEKOLLU, J. & LIEBERMAN, P. M. 2010. CTCF prevents the epigenetic drift of EBV latency promoter Qp. *PLoS Pathog,* 6**,** e1001048.

THAIN, A., JENKINS, O., CLARKE, A. R. & GASTON, K. 1996. CpG methylation directly inhibits binding of the human papillomavirus type 16 E2 protein to specific DNA sequences. *J Virol,* 70**,** 7233-5.

THEELEN, W., LITJENS, R. J., VINOKUROVA, S., HAESEVOETS, A., REIJANS, M., SIMONS, G., SMEDTS, F., HERRINGTON, C. S., RAMAEKERS, F. C., VON KNEBEL DOEBERITZ, M., SPEEL, E. J. & HOPMAN, A. H. 2013. Human papillomavirus multiplex ligation-dependent probe amplification assay for the assessment of viral load, integration, and gain of telomerase-related genes in cervical malignancies. *Hum Pathol*, 44, 2410-8

THIERRY, F. & HOWLEY, P. M. 1991. Functional analysis of E2-mediated repression of the HPV18 P105 promoter. *New Biol,* 3**,** 90-100.

THOMAS, J. T., HUBERT, W. G., RUESCH, M. N. & LAIMINS, L. A. 1999. Human papillomavirus type 31 oncoproteins E6 and E7 are required for the maintenance of episomes

during the viral life cycle in normal human keratinocytes. *Proc Natl Acad Sci U S A,* 96**,** 8449-54.

THORLAND, E. C., MYERS, S. L., GOSTOUT, B. S. & SMITH, D. I. 2003. Common fragile sites are preferential targets for HPV16 integrations in cervical tumors. *Oncogene,* 22**,** 1225-37.

THROWER, J. S., HOFFMAN, L., RECHSTEINER, M. & PICKART, C. M. 2000. Recognition of the polyubiquitin proteolytic signal. *EMBO J,* 19**,** 94-102.

TOMAIC, V., PIM, D. & BANKS, L. 2009. The stability of the human papillomavirus E6 oncoprotein is E6AP dependent. *Virology,* 393**,** 7-10.

TORRANO, V., NAVASCUES, J., DOCQUIER, F., ZHANG, R., BURKE, L. J., CHERNUKHIN, I., FARRAR, D., LEON, J., BERCIANO, M. T., RENKAWITZ, R., KLENOVA, E., LAFARGA, M. & DELGADO, M. D. 2006. Targeting of CTCF to the nucleolus inhibits nucleolar transcription through a poly(ADP-ribosyl)ation-dependent mechanism. *J Cell Sci,* 119**,** 1746-59.

TURAN, T., KALANTARI, M., CALLEJA-MACIAS, I. E., CUBIE, H. A., CUSCHIERI, K., VILLA, L. L., SKOMEDAL, H., BARRERA-SALDANA, H. A. & BERNARD, H. U. 2006. Methylation of the human papillomavirus-18 L1 gene: a biomarker of neoplastic progression? *Virology,* 349**,** 175-83.

UNDERBRINK, M. P., HOWIE, H. L., BEDARD, K. M., KOOP, J. I. & GALLOWAY, D. A. 2008. E6 proteins from multiple human betapapillomavirus types degrade Bak and protect keratinocytes from apoptosis after UVB irradiation. *J Virol,* 82**,** 10408-17.

VAILLANCOURT, P., NOTTOLI, T., CHOE, J. & BOTCHAN, M. R. 1990. The E2 transactivator of bovine papillomavirus type 1 is expressed from multiple promoters. *J Virol,* 64**,** 3927-37.

VAN DOORSLAER, K., TAN, Q., XIRASAGAR, S., BANDARU, S., GOPALAN, V., MOHAMOUD, Y., HUYEN, Y. & MCBRIDE, A. A. 2013. The Papillomavirus Episteme: a central resource for papillomavirus sequence data and analysis. *Nucleic Acids Res,* 41**,** D571-8.

VAN TINE, B. A., DAO, L. D., WU, S. Y., SONBUCHNER, T. M., LIN, B. Y., ZOU, N., CHIANG, C. M., BROKER, T. R. & CHOW, L. T. 2004a. Human papillomavirus (HPV) origin-binding protein associates with mitotic spindles to enable viral DNA partitioning. *Proc Natl Acad Sci U S A,* 101**,** 4030-5.

VAN TINE, B. A., KAPPES, J. C., BANERJEE, N. S., KNOPS, J., LAI, L., STEENBERGEN, R. D., MEIJER, C. L., SNIJDERS, P. J., CHATIS, P., BROKER, T. R., MOEN, P. T., JR. & CHOW, L. T. 2004b. Clonal selection for transcriptionally active viral oncogenes during progression to cancer. *J Virol,* 78**,** 11172-86.

VAN WIJK, S. J. & TIMMERS, H. T. 2010. The family of ubiquitin-conjugating enzymes (E2s): deciding between life and death of proteins. *FASEB J,* 24**,** 981-93.

VARGOVA, K., CURIK, N., BURDA, P., BASOVA, P., KULVAIT, V., POSPISIL, V., SAVVULIDI, F., KOKAVEC, J., NECAS, E., BERKOVA, A., OBRTLIKOVA, P., KARBAN, J., MRAZ, M., POSPISILOVA, S., MAYER, J., TRNENY, M., ZAVADIL, J. & STOPKA, T. 2011. MYB transcriptionally regulates the miR-155 host gene in chronic lymphocytic leukemia. *Blood,* 117**,** 3816-25.

VEERARAGHAVALU, K., SUBBAIAH, V. K., SRIVASTAVA, S., CHAKRABARTI, O., SYAL, R. & KRISHNA, S. 2005. Complementation of human papillomavirus type 16 E6 and E7 by Jagged1-specific Notch1-phosphatidylinositol 3-kinase signaling involves pleiotropic oncogenic functions independent of CBF1;Su(H);Lag-1 activation. *J Virol,* 79**,** 7889-98.

VETCHINOVA, A. S., AKOPOV, S. B., CHERNOV, I. P., NIKOLAEV, L. G. & SVERDLOV, E. D. 2006. Two-dimensional electrophoretic mobility shift assay: identification and

mapping of transcription factor CTCF target sequences within an FXYD5-COX7A1 region of human chromosome 19. *Anal Biochem,* 354**,** 85-93.

VINOKUROVA, S. & VON KNEBEL DOEBERITZ, M. 2011. Differential methylation of the HPV 16 upstream regulatory region during epithelial differentiation and neoplastic transformation. *PLoS One,* 6**,** e24451.

VINOKUROVA, S., WENTZENSEN, N., KRAUS, I., KLAES, R., DRIESCH, C., MELSHEIMER, P., KISSELJOV, F., DURST, M., SCHNEIDER, A. & VON KNEBEL DOEBERITZ, M. 2008. Type-dependent integration frequency of human papillomavirus genomes in cervical lesions. *Cancer Res,* 68**,** 307-13.

VOITENLEITNER, C. & BOTCHAN, M. 2002. E1 protein of bovine papillomavirus type 1 interferes with E2 protein-mediated tethering of the viral DNA to mitotic chromosomes. *J Virol,* 76**,** 3440-51.

WADA, Y., OHTA, Y., XU, M., TSUTSUMI, S., MINAMI, T., INOUE, K., KOMURA, D., KITAKAMI, J., OSHIDA, N., PAPANTONIS, A., IZUMI, A., KOBAYASHI, M., MEGURO, H., KANKI, Y., MIMURA, I., YAMAMOTO, K., MATAKI, C., HAMAKUBO, T., SHIRAHIGE, K., ABURATANI, H., KIMURA, H., KODAMA, T., COOK, P. R. & IHARA, S. 2009. A wave of nascent transcription on activated human genes. *Proc Natl Acad Sci U S A,* 106**,** 18357-61.

WALBOOMERS, J. M., JACOBS, M. V., MANOS, M. M., BOSCH, F. X., KUMMER, J. A., SHAH, K. V., SNIJDERS, P. J., PETO, J., MEIJER, C. J. & MUNOZ, N. 1999. Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. *J Pathol,* 189**,** 12-9.

WALLACE, J. A. & FELSENFELD, G. 2007. We gather together: insulators and genome organization. *Curr Opin Genet Dev,* 17**,** 400-7.

WANG, H. K., DUFFY, A. A., BROKER, T. R. & CHOW, L. T. 2009. Robust production and passaging of infectious HPV in squamous epithelium of primary human keratinocytes. *Genes Dev,* 23**,** 181-94.

WANG, L., DAI, S. Z., CHU, H. J., CUI, H. F. & XU, X. Y. 2013. Integration sites and genotype distributions of human papillomavirus in cervical intraepithelial neoplasia. *Asian Pac J Cancer Prev,* 14**,** 3837-41.

WANG, Q., GRIFFIN, H., SOUTHERN, S., JACKSON, D., MARTIN, A., MCINTOSH, P., DAVY, C., MASTERSON, P. J., WALKER, P. A., LASKEY, P., OMARY, M. B. & DOORBAR, J. 2004. Functional analysis of the human papillomavirus type 16 E1=E4 protein provides a mechanism for in vivo and in vitro keratin filament reorganization. *J Virol,* 78**,** 821-33.

WANG, X., MEYERS, C., WANG, H. K., CHOW, L. T. & ZHENG, Z. M. 2011. Construction of a full transcription map of human papillomavirus type 18 during productive viral infection. *J Virol,* 85**,** 8080-92.

WATSON, R. A., THOMAS, M., BANKS, L. & ROBERTS, S. 2003. Activity of the human papillomavirus E6 PDZ-binding motif correlates with an enhanced morphological transformation of immortalized human keratinocytes. *J Cell Sci,* 116**,** 4925-34.

WENDT, K. S., YOSHIDA, K., ITOH, T., BANDO, M., KOCH, B., SCHIRGHUBER, E., TSUTSUMI, S., NAGAE, G., ISHIHARA, K., MISHIRO, T., YAHATA, K., IMAMOTO, F., ABURATANI, H., NAKAO, M., IMAMOTO, N., MAESHIMA, K., SHIRAHIGE, K. & PETERS, J. M. 2008. Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature,* 451**,** 796-801.

WETH, O. & RENKAWITZ, R. 2011. CTCF function is modulated by neighboring DNA binding factors. *Biochem Cell Biol,* 89**,** 459-68.

WHITE, E. A., SOWA, M. E., TAN, M. J., JEUDY, S., HAYES, S. D., SANTHA, S., MUNGER, K., HARPER, J. W. & HOWLEY, P. M. 2012. Systematic identification of interactions

between host cell proteins and E7 oncoproteins from diverse human papillomaviruses. *Proc Natl Acad Sci U S A,* 109**,** E260-7.

WILSON, R., FEHRMANN, F. & LAIMINS, L. A. 2005. Role of the E1--E4 protein in the differentiation-dependent life cycle of human papillomavirus type 31. *J Virol,* 79**,** 6732-40.

WILSON, R. & LAIMINS, L. A. 2005. Differentiation of HPV-containing cells using organotypic "raft" culture or methylcellulose. *Methods Mol Med,* 119**,** 157-69.

WILSON, R., RYAN, G. B., KNIGHT, G. L., LAIMINS, L. A. & ROBERTS, S. 2007. The full-length E1E4 protein of human papillomavirus type 18 modulates differentiation-dependent viral DNA amplification and late gene expression. *Virology,* 362**,** 453-60.

WINER, R. L., HUGHES, J. P., FENG, Q., O'REILLY, S., KIVIAT, N. B., HOLMES, K. K. & KOUTSKY, L. A. 2006. Condom use and the risk of genital human papillomavirus infection in young women. *N Engl J Med,* 354**,** 2645-54.

WINER, R. L., KIVIAT, N. B., HUGHES, J. P., ADAM, D. E., LEE, S. K., KUYPERS, J. M. & KOUTSKY, L. A. 2005. Development and duration of human papillomavirus lesions, after initial infection. *J Infect Dis,* 191**,** 731-8.

WISE-DRAPER, T. M., ALLEN, H. V., JONES, E. E., HABASH, K. B., MATSUO, H. & WELLS, S. I. 2006. Apoptosis inhibition by the human DEK oncoprotein involves interference with p53 functions. *Mol Cell Biol,* 26**,** 7506-19.

WISE-DRAPER, T. M., ALLEN, H. V., THOBE, M. N., JONES, E. E., HABASH, K. B., MUNGER, K. & WELLS, S. I. 2005. The human DEK proto-oncogene is a senescence inhibitor and an upregulated target of high-risk human papillomavirus E7. *J Virol,* 79**,** 14309-17.

WOODMAN, C. B., COLLINS, S., ROLLASON, T. P., WINTER, H., BAILEY, A., YATES, M. & YOUNG, L. S. 2003. Human papillomavirus type 18 and rapidly progressing cervical intraepithelial neoplasia. *Lancet,* 361**,** 40-3.

XI, L. F., DEMERS, G. W., KOUTSKY, L. A., KIVIAT, N. B., KUYPERS, J., WATTS, D. H., HOLMES, K. K. & GALLOWAY, D. A. 1995. Analysis of human papillomavirus type 16 variants indicates establishment of persistent infection. *J Infect Dis,* 172**,** 747-55.

XIAO, T., WALLACE, J. & FELSENFELD, G. 2011. Specific sites in the C terminus of CTCF interact with the SA2 subunit of the cohesin complex and are required for cohesin-dependent insulation activity. *Mol Cell Biol,* 31**,** 2174-83.

XIE, X., MIKKELSEN, T. S., GNIRKE, A., LINDBLAD-TOH, K., KELLIS, M. & LANDER, E. S. 2007. Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proc Natl Acad Sci U S A,* 104**,** 7145-50.

XU, B., CHOTEWUTMONTRI, S., WOLF, S., KLOS, U., SCHMITZ, M., DURST, M. & SCHWARZ, E. 2013. Multiplex Identification of Human Papillomavirus 16 DNA Integration Sites in Cervical Carcinomas. *PLoS One,* 8**,** e66693.

XU, M., LUO, W., ELZI, D. J., GRANDORI, C. & GALLOWAY, D. A. 2008. NFX1 interacts with mSin3A/histone deacetylase to repress hTERT transcription in keratinocytes. *Mol Cell Biol,* 28**,** 4819-28.

YOU, J., CROYLE, J. L., NISHIMURA, A., OZATO, K. & HOWLEY, P. M. 2004. Interaction of the bovine papillomavirus E2 protein with Brd4 tethers the viral DNA to host mitotic chromosomes. *Cell,* 117**,** 349-60.

YU, T., FERBER, M. J., CHEUNG, T. H., CHUNG, T. K., WONG, Y. F. & SMITH, D. I. 2005. The role of viral integration in the development of cervical cancer. *Cancer Genet Cytogenet,* 158**,** 27-34.

YU, W., GINJALA, V., PANT, V., CHERNUKHIN, I., WHITEHEAD, J., DOCQUIER, F., FARRAR, D., TAVOOSIDANA, G., MUKHOPADHYAY, R., KANDURI, C., OSHIMURA, M., FEINBERG,

A. P., LOBANENKOV, V., KLENOVA, E. & OHLSSON, R. 2004. Poly(ADP-ribosyl)ation regulates CTCF-dependent chromatin insulation. *Nat Genet,* 36**,** 1105-10.

YUGAWA, T. & KIYONO, T. 2009. Molecular mechanisms of cervical carcinogenesis by high-risk human papillomaviruses: novel functions of E6 and E7 oncoproteins. *Rev Med Virol,* 19**,** 97-113.

YUSUFZAI, T. M., TAGAMI, H., NAKATANI, Y. & FELSENFELD, G. 2004. CTCF tethers an insulator to subnuclear sites, suggesting shared insulator mechanisms across species. *Mol Cell,* 13**,** 291-8.

ZAMPIERI, M., GUASTAFIERRO, T., CALABRESE, R., CICCARONE, F., BACALINI, M. G., REALE, A., PERILLI, M., PASSANANTI, C. & CAIAFA, P. 2012. ADP-ribose polymers localized on Ctcf-Parp1-Dnmt1 complex prevent methylation of Ctcf target sites. *Biochem J,* 441**,** 645-52.

ZANIER, K., OULD M'HAMED OULD SIDI, A., BOULADE-LADAME, C., RYBIN, V., CHAPPELLE, A., ATKINSON, A., KIEFFER, B. & TRAVE, G. 2012. Solution structure analysis of the HPV16 E6 oncoprotein reveals a self-association mechanism required for E6-mediated degradation of p53. *Structure,* 20**,** 604-17.

ZHAO, Z., TAVOOSIDANA, G., SJOLINDER, M., GONDOR, A., MARIANO, P., WANG, S., KANDURI, C., LEZCANO, M., SANDHU, K. S., SINGH, U., PANT, V., TIWARI, V., KURUKUTI, S. & OHLSSON, R. 2006. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat Genet,* 38**,** 1341-7.

ZIEBARTH, J. D., BHATTACHARYA, A. & CUI, Y. 2013. CTCFBSDB 2.0: a database for CTCF-binding sites and genome organization. *Nucleic Acids Res,* 41**,** D188-94.

ZIEGERT, C., WENTZENSEN, N., VINOKUROVA, S., KISSELJOV, F., EINENKEL, J., HOECKEL, M. & VON KNEBEL DOEBERITZ, M. 2003. A comprehensive analysis of HPV integration loci in anogenital lesions combining transcript and genome-based amplification techniques. *Oncogene,* 22**,** 3977-84.

ZIELKE, K., FULL, F., TEUFERT, N., SCHMIDT, M., MULLER-FLECKENSTEIN, I., ALBERTER, B. & ENSSER, A. 2012. The insulator protein CTCF binding sites in the orf73/LANA promoter region of herpesvirus saimiri are involved in conferring episomal stability in latently infected human T cells. *J Virol,* 86**,** 1862-73.

ZIMBER-STROBL, U., STROBL, L. J., MEITINGER, C., HINRICHS, R., SAKAI, T., FURUKAWA, T., HONJO, T. & BORNKAMM, G. W. 1994. Epstein-Barr virus nuclear antigen 2 exerts its transactivating function through interaction with recombination signal binding protein RBP-J kappa, the homologue of Drosophila Suppressor of Hairless. *EMBO J,* 13**,** 4973-82.

ZLATANOVA, J. & CAIAFA, P. 2009a. CCCTC-binding factor: to loop or to bridge. *Cell Mol Life Sci,* 66**,** 1647-60.

ZLATANOVA, J. & CAIAFA, P. 2009b. CTCF and its protein partners: divide and rule? *J Cell Sci,* 122**,** 1275-84.

ZUR HAUSEN, H. 2009. Papillomaviruses in the causation of human cancers - a brief historical account. *Virology,* 384**,** 260-5.

6 Appendix

6.1 Expression constructs for human CTCF

6.1.1 Amino acid sequence of expressed product

Histidine tag

CTCF protein

MGHHHHHHHHHHSSGHIEGRHMEGDAVEAIVEESETFIKGKERKTYQRRREGGQEEDACH 60
LPQNQTDGGEVVQDVNSSVQMVMMEQLDPTLLQMKTEVMEGTVAPEAEAAVDDTQIITLQ 120
VVNMEEQPINIGELQLVQVPVPVTVPVATTSVEELQGAYENEVSKEGLAESEPMICHTLP 180
LPEGFQVVKVGANGEVETLEQGELPPQEDPSWQKDPDYQPPAKKTKKTKKSKLRYTEEGK 240
DVDVSVYDFEEEQQEGLLSEVNAEKVVGNMKPPKPTKIKKKGVKKTFQCELCSYTCPRRS 300
NLDRHMKSHTDERPHKCHLCGRAFRTVTLLRNHLNTHTGTRPHKCPDCDMAFVTSGELVR 360
HRRYKHTHEKPFKCSMCDYASVEVSKLKRHIRSHTGERPFQCSLCSYASRDTYKLKRHMR 420
THSGEKPYECYICHARFTQSGTMKMHILQKHTENVAKFHCPHCDTVIARKSDLGVHLRKQ 480
HSYIEQGKKCRYCDAVFHERYALIQHQKSHKNEKRFKCDQCDYACRQERHMIMHKRTHTG 540
EKPYACSHCDKTFRQKQLLDMHFKRYHDPNFVPAAFVCSKCGKTFTRRNTMARHADNCAG 600
PDGVEGENGGETKKSKRGRKRKMRSKKEDSSDSENAEPDLDDNEDEEEPAVEIEPEPEPQ 660
PVTPAPPPAKKRRGRPPGRTNQPKQNQPTAIIQVEDQNTGAIENIIVEVKKEPDAEPAEG 720
EEEEAQPAATDAPNGDLTPEMILSMMDR 748

6.1.2 pCi-T7-hCTCF-HIS-tag construct

Restriction site

Remaining sequence of previous vector from which the CTCF construct was excised

CTCF protein sequence

T7 promoter

pCi sequence containing the insert for CTCF expression

```
TCAATATTGGCCATTAGCCATATTATTCATTGGTTATATAGCATAAATCAATATTGGCTA 60
TTGGCCATTGCATACGTTGTATCTATATCATAATATGTACATTTATATTGGCTCATGTCC 120
AATATGACCGCCATGTTGGCATTGATTATTGACTAGTTATTAATAGTAATCAATTACGGG 180
GTCATTAGTTCATAGCCCATATATGGAGTTCCGCGTTACATAACTTACGGTAAATGGCCC 240
GCCTGGCTGACCGCCCAACGACCCCCGCCCATTGACGTCAATAATGACGTATGTTCCCAT 300
AGTAACGCCAATAGGGACTTTCCATTGACGTCAATGGGTGGAGTATTTACGGTAAACTGC 360
CCACTTGGCAGTACATCAAGTGTATCATATGCCAAGTCCGCCCCCTATTGACGTCAATGA 420
CGGTAAATGGCCCGCCTGGCATTATGCCCAGTACATGACCTTACGGGACTTTCCTACTTG 480
GCAGTACATCTACGTATTAGTCATCGCTATTACCATGGTGATGCGGTTTTGGCAGTACAC 540
CAATGGGCGTGGATAGCGGTTTGACTCACGGGGATTTCCAAGTCTCCACCCCATTGACGT 600
CAATGGGAGTTTGTTTTGGCACCAAAATCAACGGGACTTTCCAAAATGTCGTAATAACCC 660
CGCCCCGTTGACGCAAATGGGCGGTAGGCGTGTACGGTGGGAGGTCTATATAAGCAGAGC 720
TCGTTTAGTGAACCGTCAGATCACTAGAAGCTTTATTGCGGTAGTTTATCACAGTTAAAT 780
TGCTAACGCAGTCAGTGCTTCTGACACAACAGTCTCGAACTTAAGCTGCAGAAGTTGGTC 840
GTGAGGCACTGGGCAGGTAAGTATCAAGGTTACAAGACAGGTTTAAGGAGACCAATAGAA 900
ACTGGGCTTGTCGAGACAGAGAAGACTCTTGCGTTTCTGATAGGCACCTATTGGTCTTAC 960
TGACATCCACTTTGCCTTTCTCTCCACAGGTGTCCACTCCCAGTTCAATTACAGCTCTTA 1020
AGGCTAGAGTACTTAATACGACTCACTATAGGCTAGAAATAATTTTGTTTAACTTTAAGA 1080
AGGAGATATACCATGGGCCATCATCATCATCATCATCATCATCATCACAGCAGCGGCCAT 1140
ATCGAAGGTCGTCATATGGAAGGTGATGCAGTCGAAGCCATTGTGGAGGAGTCCGAAACT 1200
TTTATTAAAGGAAAGGAGAGAAAGACTTACCAGAGACGCCGGGAAGGGGGCCAGGAAGAA 1260
GATGCCTGCCACTTACCCCAGAACCAGACGGATGGGGGTGAGGTGGTCCAGGATGTCAAC 1320
AGCAGTGTACAGATGGTGATGATGGAACAGCTGGACCCCACCCTTCTTCAGATGAAGACT 1380
GAAGTAATGGAGGGCACAGTGGCTCCAGAAGCAGAGGCTGCTGTGGACGATACCCAGATT 1440
ATAACTTTACAGGTTGTAAATATGGAGGAACAGCCCATAAACATAGGAGAACTTCAGCTT 1500
GTTCAAGTACCTGTTCCTGTGACTGTACCTGTTGCTACCACTTCAGTAGAAGAACTTCAG 1560
GGGGCTTATGAAAATGAAGTGTCTAAAGAGGGCCTTGCGGAAAGTGAACCCATGATATGC 1620
CACACCCTACCTTTGCCTGAAGGGTTTCAGGTGGTTAAAGTGGGGGCCAATGGAGAGGTG 1680
GAGACACTAGAACAAGGGGAACTTCCACCCCAGGAAGATCCTAGTTGGCAAAAAGACCCA 1740
GACTATCAGCCACCAGCCAAAAAAACAAAGAAAACCAAAAAGAGCAAACTGCGTTATACA 1800
GAGGAGGGCAAAGATGTAGATGTGTCTGTCTACGATTTTGAGGAAGAACAGCAGGAGGGT 1860
CTGCTATCAGAGGTTAATGCAGAGAAAGTGGTTGGTAATATGAAGCCTCCAAAGCCAACA 1920
AAAATTAAAAAGAAAGGTGTAAAGAAGACATTCCAGTGTGAGCTTTGCAGTTACACGTGT 1980
CCACGGCGTTCAAATTTGGATCGTCACATGAAAAGCCACACTGATGAGAGACCACACAAG 2040
TGCCATCTCTGTGGCAGGGCATTCAGAACAGTCACCCTCCTGAGGAATCACCTTAACACA 2100
CACACAGGTACTCGTCCTCACAAGTGCCCAGACTGCGACATGGCCTTTGTGACCAGTGGA 2160
GAATTGGTTCGGCATCGTCGTTACAAACACACCCACGAGAAGCCATTCAAGTGTTCCATG 2220
TGCGATTACGCCAGTGTAGAAGTCAGCAAATTAAAACGTCACATTCGCTCTCATACTGGA 2280
GAGCGTCCGTTTCAGTGCAGTTTGTGCAGTTATGCCAGCAGGGACACATACAAGCTGAAA 2340
AGGCACATGAGAACCCATTCAGGGGAAAAGCCTTATGAATGTTATATTTGTCATGCTCGG 2400
```

```
TTTACCCAAAGTGGTACCATGAAGATGCACATTTTACAGAAGCACACAGAAAATGTGGCC 2460
AAATTTCACTGTCCCCACTGTGACACAGTCATAGCCCGAAAAAGTGATTTGGGTGTCCAC 2520
TTGCGAAAGCAGCATTCCTATATTGAGCAAGGCAAGAAATGCCGTTACTGTGATGCTGTG 2580
TTTCATGAGCGCTATGCCCTCATCCAGCATCAGAAGTCACACAAGAATGAGAAGCGCTTT 2640
AAGTGTGACCAGTGTGATTACGCTTGTAGACAGGAGAGGCACATGATCATGCACAAGCGC 2700
ACCCACACCGGGGAGAAGCCTTACGCCTGCAGCCACTGCGATAAGACCTTCCGCCAGAAG 2760
CAGCTTCTCGACATGCACTTCAAGCGCTATCACGACCCCAACTTCGTCCCTGCGGCTTTT 2820
GTCTGTTCTAAGTGTGGGAAAACATTTACACGTCGGAATACCATGGCAAGACATGCTGAT 2880
AATTGTGCTGGCCCAGATGGCGTAGAGGGGGAAAATGGAGGAGAAACGAAGAAGAGTAAA 2940
CGTGGAAGAAAAAGAAAGATGCGCTCTAAGAAAGAAGATTCCTCTGACAGTGAAAATGCT 3000
GAACCAGATCTGGACGACAATGAGGATGAGGAGGAGCCTGCCGTAGAAATTGAACCTGAG 3060
CCAGAGCCTCAGCCTGTGACCCCAGCCCCACCACCCGCCAAGAAGCGGAGAGGACGACCC 3120
CCTGGCAGAACCAACCAGCCCAAACAGAACCAGCCAACAGCTATCATTCAGGTTGAAGAC 3180
CAGAATACAGGTGCAATTGAGAACATTATAGTTGAAGTAAAAAAAGAGCCAGATGCTGAG 3240
CCCGCAGAGGGAGAGGAAGAGGAGGCCCAGCCAGCTGCCACAGATGCCCCCAACGGAGAC 3300
CTCACGCCCGAGATGATCCTCAGCATGATGGACCGGTGATGGCGGAGCCTTGTGCGTCGC 3360
CAGGACTTCTCTGGGCTGTGTTTAAACGGCCCGCATCTTAATTTTTCTCCCTTCTTTCTT 3420
TTTTTGGCTTTGGGAAAAGCATCATTTTACCAAACATACCGAGAACGAAAACTTCAAGGA 3480
TGATGTTAGAAAAAAATGTGATTTAACTAGAACTTGCTGTCTGATGTTAGCAAATCATGG 3540
AATGTTCTGAGTCCCTGAGGGTTTACTGTGAAGTGCTGAGGACAGTGTTGACAACTAACT 3600
CGTTTTCCTAGATGGAAACGGAGACATTGACCCCTCCCTCCATGTGGTAAACCACTCCAG 3660
AATGGCCGCACTCGAGAATTCACGCGTGGTACCTCTAGAGTCGACCCGGGCGGCCGCTTC 3720
GAGCAGACATGATAAGATACATTGATGAGTTTGGACAAACCACAACTAGAATGCAGTGAA 3780
AAAAATGCTTTATTTGTGAAATTTGTGATGCTATTGCTTTATTTGTAACCATTATAAGCT 3840
GCAATAAACAAGTTAACAACAACAATTGCATTCATTTTATGTTTCAGGTTCAGGGGGAGA 3900
TGTGGGAGGTTTTTTAAAGCAAGTAAAACCTCTACAAATGTGGTAAAATCGATAAGGATC 3960
CGGGCTGGCGTAATAGCGAAGAGGCCCGCACCGATCGCCCTTCCCAACAGTTGCGCAGCC 4020
TGAATGGCGAATGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTAC 4080
GCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCC 4140
TTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCGTCAAGCTCTAAATCGGGGGCTCCCTTT 4200
AGGGTTCCGATTTAGTGCTTTACGGCACCTCGACCCCAAAAAACTTGATTAGGGTGATGG 4260
TTCACGTAGTGGGCCATCGCCCTGATAGACGGTTTTTCGCCCTTTGACGTTGGAGTCCAC 4320
GTTCTTTAATAGTGGACTCTTGTTCCAAACTGGAACAACACTCAACCCTATCTCGGTCTA 4380
TTCTTTTGATTTATAAGGGATTTTGCCGATTTCGGCCTATTGGTTAAAAAATGAGCTGAT 4440
TTAACAAAAATTTAACGCGAATTTTAACAAAATATTAACGCTTACAATTTCCTGATGCGG 4500
TATTTTCTCCTTACGCATCTGTGCGGTATTTCACACCGCATATGGTGCACTCTCAGTACA 4560
ATCTGCTCTGATGCCGCATAGTTAAGCCAGCCCCGACACCCGCCAACACCCGCTGACGCG 4620
CCCTGACGGGCTTGTCTGCTCCCGGCATCCGCTTACAGACAAGCTGTGACCGTCTCCGGG 4680
AGCTGCATGTGTCAGAGGTTTTCACCGTCATCACCGAAACGCGCGAGACGAAAGGGCCTC 4740
GTGATACGCCTATTTTTATAGGTTAATGTCATGATAATAATGGTTTCTTAGACGTCAGGT 4800
GGCACTTTTCGGGGAAATGTGCGCGGAACCCCTATTTGTTTATTTTTCTAAATACATTCA 4860
AATATGTATCCGCTCATGAGACAATAACCCTGATAAATGCTTCAATAATATTGAAAAAGG 4920
AAGAGTATGAGTATTCAACATTTCCGTGTCGCCCTTATTCCCTTTTTTGCGGCATTTTGC 4980
CTTCCTGTTTTTGCTCACCCAGAAACGCTGGTGAAAGTAAAAGATGCTGAAGATCAGTTG 5040
GGTGCACGAGTGGGTTACATCGAACTGGATCTCAACAGCGGTAAGATCCTTGAGAGTTTT 5100
CGCCCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTA 5160
TTATCCCGTATTGACGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAAT 5220
GACTTGGTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGA 5280
GAATTATGCAGTGCTGCCATAACCATGAGTGATAACACTGCGGCCAACTTACTTCTGACA 5340
ACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATGGGGGATCATGTAACT 5400
CGCCTTGATCGTTGGGAACCGGAGCTGAATGAAGCCATACCAAACGACGAGCGTGACACC 5460
```

```
ACGATGCCTGTAGCAATGGCAACAACGTTGCGCAAACTATTAACTGGCGAACTACTTACT 5520
CTAGCTTCCCGGCAACAATTAATAGACTGGATGGAGGCGGATAAAGTTGCAGGACCACTT 5580
CTGCGCTCGGCCCTTCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGT 5640
GGGTCTCGCGGTATCATTGCAGCACTGGGGCCAGATGGTAAGCCCTCCCGTATCGTAGTT 5700
ATCTACACGACGGGGAGTCAGGCAACTATGGATGAACGAAATAGACAGATCGCTGAGATA 5760
GGTGCCTCACTGATTAAGCATTGGTAACTGTCAGACCAAGTTTACTCATATATACTTTAG 5820
ATTGATTTAAAACTTCATTTTTAATTTAAAAGGATCTAGGTGAAGATCCTTTTTGATAAT 5880
CTCATGACCAAAATCCCTTAACGTGAGTTTTCGTTCCACTGAGCGTCAGACCCCGTAGAA 5940
AAGATCAAAGGATCTTCTTGAGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACA 6000
AAAAAACCACCGCTACCAGCGGTGGTTTGTTTGCCGGATCAAGAGCTACCAACTCTTTTT 6060
CCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTTCTTCTAGTGTAGCCG 6120
TAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTGCTAATC 6180
CTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGA 6240
CGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCGTGCACACAGCCC 6300
AGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAGAAAGC 6360
GCCACGCTTCCCGAAGGGAGAAAGGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAACA 6420
GGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGGTATCTTTATAGTCCTGTCGGG 6480
TTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTA 6540
TGGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGCCTTTTGCT 6600
CACATGGCTCGACAGATCT
```

6.1.3 pDrive-Sp6-hCTCF-HIS-tag construct

Restriction site

Remaining sequence of previous vector from which the CTCF construct was excised

CTCF protein sequence

SP6 promoter

pDrive sequence containing the inverted insert for CTCF expression

```
GCGCCCAATACGCAAACCGCCTCTCCCCGCGCGTTGGCCGATTCATTAATGCAGCTGGCA      60
CGACAGGTTTCCCGACTGGAAAGCGGGCAGTGAGCGCAACGCAATTAATGTGAGTTAGCT     120
CACTCATTAGGCACCCCAGGCTTTACACTTTATGCTTCCGGCTCGTATGTTGTGTGGAAT     180
TGTGAGCGGATAACAATTTCACACAGGAAACAGCTATGACCATGATTACGCCAAGCTCTA     240
ATACGACTCACTATAGGGAAAGCTCGGTACCACGCATGCTGCAGACGCGTTACGTATCGG     300
ATCCAGAATTCTCGAGTGCGGCCATTCTGGAGTGGTTTACCACATGGAGGGAGGGGTCAA     360
TGTCTCCGTTTCCATCTAGGAAAACGAGTTAGTTGTCAACACTGTCCTCAGCACTTCACA     420
GTAAACCCTCAGGGACTCAGAACATTCCATGATTTGCTAACATCAGACAGCAAGTTCTAG     480
TTAAATCACATTTTTTTCTAACATCATCCTTGAAGTTTTCGTTCTCGGTATGTTTGGTAA     540
AATGATGCTTTTCCCAAAGCCAAAAAAAGAAAGAAGGGAGAAAAATTAAGATGCGGGCCG     600
TTTAAACACAGCCCAGAGAAGTCCTGGCGACGCACAAGGCTCCGCCATCACCGGTCCATC     660
ATGCTGAGGATCATCTCGGGCGTGAGGTCTCCGTTGGGGGCATCTGTGGCAGCTGGCTGG     720
GCCTCCTCTTCCTCTCCCTCTGCGGGCTCAGCATCTGGCTCTTTTTTTACTTCAACTATA     780
ATGTTCTCAATTGCACCTGTATTCTGGTCTTCAACCTGAATGATAGCTGTTGGCTGGTTC     840
TGTTTGGGCTGGTTGGTTCTGCCAGGGGGTCGTCCTCTCCGCTTCTTGGCGGGTGGTGGG     900
GCTGGGGTCACAGGCTGAGGCTCTGGCTCAGGTTCAATTTCTACGGCAGGCTCCTCCTCA     960
TCCTCATTGTCGTCCAGATCTGGTTCAGCATTTTCACTGTCAGAGGAATCTTCTTTCTTA    1020
GAGCGCATCTTTCTTTTTCTTCCACGTTTACTCTTCTTCGTTTCTCCTCCATTTTCCCCC    1080
TCTACGCCATCTGGGCCAGCACAATTATCAGCATGTCTTGCCATGGTATTCCGACGTGTA    1140
AATGTTTTCCCACACTTAGAACAGACAAAAGCCGCAGGGACGAAGTTGGGGTCGTGATAG    1200
CGCTTGAAGTGCATGTCGAGAAGCTGCTTCTGGCGGAAGGTCTTATCGCAGTGGCTGCAG    1260
GCGTAAGGCTTCTCCCCGGTGTGGGTGCGCTTGTGCATGATCATGTGCCTCTCCTGTCTA    1320
CAAGCGTAATCACACTGGTCACACTTAAAGCGCTTCTCATTCTTGTGTGACTTCTGATGC    1380
TGGATGAGGGCATAGCGCTCATGAAACACAGCATCACAGTAACGGCATTTCTTGCCTTGC    1440
TCAATATAGGAATGCTGCTTTCGCAAGTGGACACCCAAATCACTTTTTCGGGCTATGACT    1500
GTGTCACAGTGGGGACAGTGAAATTTGGCCACATTTTCTGTGTGCTTCTGTAAAATGTGC    1560
ATCTTCATGGTACCACTTTGGGTAAACCGAGCATGACAAATATAACATTCATAAGGCTTT    1620
TCCCCTGAATGGGTTCTCATGTGCCTTTTCAGCTTGTATGTGTCCCTGCTGGCATAACTG    1680
CACAAACTGCACTGAAACGGACGCTCTCCAGTATGAGAGCGAATGTGACGTTTTAATTTG    1740
CTGACTTCTACACTGGCGTAATCGCACATGGAACACTTGAATGGCTTCTCGTGGGTGTGT    1800
TTGTAACGACGATGCCGAACCAATTCTCCACTGGTCACAAAGGCCATGTCGCAGTCTGGG    1860
CACTTGTGAGGACGAGTACCTGTGTGTGTGTTAAGGTGATTCCTCAGGAGGGTGACTGTT    1920
CTGAATGCCCTGCCACAGAGATGGCACTTGTGTGGTCTCTCATCAGTGTGGCTTTTCATG    1980
TGACGATCCAAATTTGAACGCCGTGGACACGTGTAACTGCAAAGCTCACACTGGAATGTC    2040
TTCTTTACACCTTTCTTTTTAATTTTTGTTGGCTTTGGAGGCTTCATATTACCAACCACT    2100
TTCTCTGCATTAACCTCTGATAGCAGACCCTCCTGCTGTTCTTCCTCAAAATCGTAGACA    2160
GACACATCTACATCTTTGCCCTCCTCTGTATAACGCAGTTTGCTCTTTTTGGTTTTCTTT    2220
GTTTTTTTGGCTGGTGGCTGATAGTCTGGGTCTTTTTGCCAACTAGGATCTTCCTGGGGT    2280
GGAAGTTCCCCTTGTTCTAGTGTCTCCACCTCTCCATTGGCCCCCACTTTAACCACCTGA    2340
AACCCTTCAGGCAAAGGTAGGGTGTGGCATATCATGGGTTCACTTTCCGCAAGGCCCTCT    2400
```

```
TTAGACACTTCATTTTCATAAGCCCCCTGAAGTTCTTCTACTGAAGTGGTAGCAACAGGT 2460
ACAGTCACAGGAACAGGTACTTGAACAAGCTGAAGTTCTCCTATGTTTATGGGCTGTTCC 2520
TCCATATTTACAACCTGTAAAGTTATAATCTGGGTATCGTCCACAGCAGCCTCTGCTTCT 2580
GGAGCCACTGTGCCCTCCATTACTTCAGTCTTCATCTGAAGAAGGGTGGGGTCCAGCTGT 2640
TCCATCATCACCATCTGTACACTGCTGTTGACATCCTGGACCACCTCACCCCCATCCGTC 2700
TGGTTCTGGGGTAAGTGGCAGGCATCTTCTTCCTGGCCCCCTTCCCGGCGTCTCTGGTAA 2760
GTCTTTCTCTCCTTTCCTTTAATAAAAGTTTCGGACTCCTCCACAATGGCTTCGACTGCA 2820
TCACCTTCCATATGACGACCTTCGATATGGCCGCTGCTGTGATGATGATGATGATGATGA 2880
TGATGATGGCCCATGGTATATCTCCTTCTTAAAGTTAAACAAAATTATTTCTAGCCTATA 2940
GTGAGTCGTATTAAGTACTCTAGCCTTAAGAGCTGTAATTGAACTGGGAGTGGACACCTG 3000
TGGAGAGAAAGGCAAAGTGGATGTCAGTAAGACCAATAGGTGCCTATCAGAAACGCAAGA 3060
GTCTTCTCTGTCTCGACAAGCCCAGTTTCTATTGGTCTCCTTAAACCTGTCTTGTAACCT 3120
TGATACTTACCTGCCCAGTGCCTCACGACCAACTTCTGCAGCTTAAGTTCGAGACTGTTG 3180
TGTCAGAAGCACTGACTGCGTTAGCAATTTAACTGTGATAAACTACCGCAATAAAGCTTC 3240
TAGTGATCTGACGGTTCACTAAACGAGCTCGCGGCCGCTGTATTCTATAGTGTCACCTAA 3300
ATGGCCGCACAATTCACTGGCCGTCGTTTTACAACGTCGTGACTGGGAAAACCCTGGCGT 3360
TACCCAACTTAATCGCCTTGCAGCACATCCCCCTTTCGCCAGCTGGCGTAATAGCGAAGA 3420
GGCCCGCACCGATCGCCCTTCCCAACAGTTGCGCAGCCTGAATGGCGAATGGAAATTGTA 3480
AGCGTTAATATTTTGTTAAAATTCGCGTTAAATTTTTGTTAAATCAGCTCATTTTTTAAC 3540
CAATAGGCCGAAATCGGCAAAATCCCTTATAAATCAAAAGAATAGACCGAGATAGGGTTG 3600
AGTGTTGTTCCAGTTTGGAACAAGAGTCCACTATTAAAGAACGTGGACTCCAACGTCAAA 3660
GGGCGAAAAACCGTCTATCAGGGCGATGGCCCACTACGTGAACCATCACCCTAATCAAGT 3720
TTTTTGGGGTCGAGGTGCCGTAAAGCACTAAATCGGAACCCTAAAGGGAGCCCCCGATTT 3780
AGAGCTTGACGGGGAAAGCCGGCGAACGTGGCGAGAAAGGAAGGGAAGAAAGCGAAAGGA 3840
GCGGGCGCTAGGGCGCTGGCAAGTGTAGCGGTCACGCTGCGCGTAACCACCACACCCGCC 3900
GCGCTTAATGCGCCGCTACAGGGCGCGTCAGGTGGCACTTTTCGGGGAAATGTGCGCGGA 3960
ACCCCTATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAGACAATAA 4020
CCCTGATAAATGCTTCAATAATATTGAAAAAGGAAGAGTATGAGTATTCAACATTTCCGT 4080
GTCGCCCTTATTCCCTTTTTTGCGGCATTTTGCCTTCCTGTTTTTGCTCACCCAGAAACG 4140
CTGGTGAAAGTAAAAGATGCTGAAGATCAGTTGGGTGCACGAGTGGGTTACATCGAACTG 4200
GATCTCAACAGCGGTAAGATCCTTGAGAGTTTTCGCCCCGAAGAACGTTTTCCAATGATG 4260
AGCACTTTTAAAGTTCTGCTATGTGGCGCGGTATTATCCCGTATTGACGCCGGGCAAGAG 4320
CAACTCGGTCGCCGCATACACTATTCTCAGAATGACTTGGTTGAGTACTCACCAGTCACA 4380
GAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTATGCAGTGCTGCCATAACCATG 4440
AGTGATAACACTGCGGCCAACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACC 4500
GCTTTTTTGCACAACATGGGGGATCATGTAACTCGCCTTGATCGTTGGGAACCGGAGCTG 4560
AATGAAGCCATACCAAACGACGAGCGTGACACCACGATGCCTGTAGCAATGGCAACAACG 4620
TTGCGCAAACTATTAACTGGCGAACTACTTACTCTAGCTTCCCGGCAACAATTAATAGAC 4680
TGGATGGAGGCGGATAAAGTTGCAGGACCACTTCTGCGCTCGGCCCTTCCGGCTGGCTGG 4740
TTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGGTCTCGCGGTATCATTGCAGCACTG 4800
GGGCCAGATGGTAAGCCCTCCCGTATCGTAGTTATCTACACGACGGGGAGTCAGGCAACT 4860
ATGGATGAACGAAATAGACAGATCGCTGAGATAGGTGCCTCACTGATTAAGCATTGGTAA 4920
CTGTCAGACCAAGTTTACTCATATATACTTTAGATTGATTTAAAACTTCATTTTTAATTT 4980
AAAAGGATCTAGGTGAAGATCCTTTTTGATAATCTCATGAACAATAAAACTGTCTGCTTA 5040
CATAAACAGTAATACAAGGGGTGTTATGAGCCATATTCAACGGGAAACGTCTTGCTCTAG 5100
GCCGCGATTAAATTCCAACATGGATGCTGATTTATATGGGTATAAATGGGCTCGCGATAA 5160
TGTCGGGCAATCAGGTGCGACAATCTATCGATTGTATGGGAAGCCCGATGCGCCAGAGTT 5220
GTTTCTGAAACATGGCAAAGGTAGCGTTGCCAATGATGTTACAGATGAGATGGTCAGACT 5280
AAACTGGCTGACGGAATTTATGCCTCTTCCGACCATCAAGCATTTTATCCGTACTCCTGA 5340
TGATGCATGGTTACTCACCACTGCGATCCCCGGGAAAACAGCATTCCAGGTATTAGAAGA 5400
ATATCCTGATTCAGGTGAAAATATTGTTGATGCGCTGGCAGTGTTCCTGCGCCGGTTGCA 5460
```

```
TTCGATTCCTGTTTGTAATTGTCCTTTTAACAGCGATCGCGTATTTCGTCTCGCTCAGGC    5520
GCAATCACGAATGAATAACGGTTTGGTTGATGCGAGTGATTTTGATGACGAGCGTAATGG    5580
CTGGCCTGTTGAACAAGTCTGGAAAGAAATGCATAAACTTTTGCCATTCTCACCGGATTC    5640
AGTCGTCACTCATGGTGATTTCTCACTTGATAACCTTATTTTTGACGAGGGGAAATTAAT    5700
AGGTTGTATTGATGTTGGACGAGTCGGAATCGCAGACCGATACCAGGATCTTGCCATCCT    5760
ATGGAACTGCCTCGGTGAGTTTTCTCCTTCATTACAGAAACGGCTTTTTCAAAAATATGG    5820
TATTGATAATCCTGATATGAATAAATTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTA    5880
AGAATTAATTCATGACCAAAATCCCTTAACGTGAGTTTTCGTTCCACTGAGCGTCAGACC    5940
CCGTAGAAAAGATCAAAGGATCTTCTTGAGATCCTTTTTTTCTGCGCGTAATCTGCTGCT    6000
TGCAAACAAAAAAACCACCGCTACCAGCGGTGGTTTGTTTGCCGGATCAAGAGCTACCAA    6060
CTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTCCTTCTAG    6120
TGTAGCCGTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTC    6180
TGCTAATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGG    6240
ACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCGTGCA    6300
CACAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTGAGCTAT    6360
GAGAAAGCGCCACGCTTCCCGAAGGGAGAAAGGCGGACAGGTATCCGGTAAGCGGCAGGG    6420
TCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGGTATCTTTATAGTC    6480
CTGTCGGGTTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGC    6540
GGAGCCTATGGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGC    6600
CTTTTGCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCG    6660
CCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGAACGACCGAGCGCAGCGAGTCAGTGA    6720
GCGAGGAAGCGGAAGA
```

6.2 Sequence alignment of PCR fragments across mutation site 3000 from all four cell lines

Georgie cells

```
HPV 18 AY262282.1   CATACAGACATTAAACCACCAGGTGGTGCCAGCCTATAACATTTCAAAAAGTAA
Georgie WT         CATACAGACATTAAACCACCAGGTGGTGCCAGCCTATAACATTTCAAAAAGTAA
C3 mutant sequence CATACAGACATTAAATCACCAGGTAGTGCCAGCCTATAACATCTCAAAAAGTAA
Georgie C3         CATACAGACATTAAATCACCAGGTAGTGCCAGCCTATAACATCTCAAAAAGTAA
                   *************** ******** ***************** ***********
```

Clonetics cells

```
HPV 18 AY262282.1   CATACAGACATTAAACCACCAGGTGGTGCCAGCCTATAACATTTCAAAAAGTAA
Clonetics WT       CATACAGACATTAAACCACCAGGTGGTGCCAGCCTATAACATTTCAAAAAGTAA
C3 mutant sequence CATACAGACATTAAATCACCAGGTAGTGCCAGCCTATAACATCTCAAAAAGTAA
Clonetics C3       CATACAGACATTAAATCACCAGGTAGTGCCAGCCTATAACATCTCAAAAAGTAA
                   *************** ******** ***************** ***********
```

6.3 Primers for EMSA testing of CTCF binding sites

Template	Start	End	Fragment name	Sequence FW 5' to 3' with M13 (-21) tag at 5'	Tm FW	Sequence RV 5' to 3'	Tm RV	CTCF binding
BPV1	55	223	LCR P9	M13-aactataaaaagctgctgacagacc	59,45	agccttcccgaattacaaca	59,57	none
BPV1	580	745	Essex-1 663	M13-agcacctgatgcacctgatt	60,69	tcagctcttttccgcatttc	60,47	none
BPV1	772	950	P890 region	M13-gtccctgcttggatttgaac	59,53	cccgggttcctcattctc	60,41	weak
BPV1	843	983	Random Frag 7	M13-agcgtcatggcaaacgat	60,23	ccgatcagattccacagaca	59,63	none
BPV1	1031	1201	E2 BS no. 14 at 1125	M13-gggaaatcacctggaggtct	60,31	tcagcaaataatcttcgctttg	59,52	none
BPV1	1137	1309	Essex-3 1227	M13-tccgaagcatctgaaactcc	60,34	acaagctgcagatgtagatgactaa	59,53	two weak bands
BPV1	2327	2488	Essex+2 2427, Essex-2 2397	M13-aaatgcattggatggctacc	59,79	ggctgctcaaagcgaaag	59,82	weak
BPV1	2650	2819	Random Frag 4	M13-caaatgcagttgattgagaaaagt	59,71	tgcttggctctctcttgaca	59,86	none
BPV1	2704	2915	No CTCF 2704-2915	M13-tggactgctgttagaactgagaac	59,99	atatatcggtcccagcttgtgt	59,75	none
BPV1	2785	2984	CTCF?A 2785 - 2984	M13-ccacactctgtagtttgtcaagaga	59,9	cttgcatttccatcaaactcc	59,56	none
BPV1	2831	3027	CTCF?B 2840-3020	M13-tgcagttgtctttgcaggag	60,18	gcgcatgtacaaattgctgt	59,76	strong
BPV1	2917	3096	E+1 3025 (at 5' edge)	M13-tcagaacctaaacggtgcttt	58,89	ggccatggtgcagtagtagag	59,77	strong
BPV1	2947	3120	E+1 3025 (in middle)	M13-gccagggtggtagaggtg	59,05	agaatagtaaatgcgtccagcac	59,71	none
BPV1	2995	3174	E+1 3025 (at 3' edge)	M13-tggtacactgtctacagcaatttgt	60,04	tcttacagagtaatgccctgttgt	59,26	none
BPV1	3072	3238	Random Frag 5	M13-tgggctctactactgcaccat	59,77	cgtctgggcgatctctaaaa	60,34	none
BPV1	3576	3775	C-1 3671	M13-aggagggtcatgctttgct	59,82	gaaagtcttgcctttgacttgg	60,27	weak
BPV1	4049	4226	Essex-6 4144	M13-tgttgtgtggatttgatttgttt	59,28	tcataggcactggcacgtt	60,28	none
BPV1	4145	4354	pred.4257	M13-ccctgctcagattttatatggtt	58.60	ttcctagccctcctaagtagattg	59.36	none
BPV1	4442	4611	Essex+3 4534	M13-catcaataggatccagagctgtaa	59,65	tgaatctgcaggaacagcat	59,4	none
BPV1	4689	4863	Essex-5 4775	M13-ggacatagcggttcttgagc	59,84	tcctcctgtatcccctaaacc	59,29	medium
BPV1	4910	5051	Neg. Contr.	M13-cccgcagtattgcctctaaa	60,22	agcacagctggttctgcttc	60,74	none

Table 40) Primers for EMSA testing of BPV1 part I

Template	Start	End	Fragment name	Sequence FW 5' to 3' with M13 (-21) tag at 5'	Tm FW	Sequence RV 5' to 3'	Tm RV	CTCF binding
BPV1	5105	5263	Essex-4 5190	M13-ttacaacacgtagcgggaca	60,17	cctcaaaacctgcttgctct	59,62	none
BPV1	5232	5422	pred. 5316	M13-acccttgcatgaagagcaa	59.38	aagtgtcaggtgagccatagg	59.22	none
BPV1	5475	5639	Random Frag 2	M13-cgatgacactactactacaccaatca	59,55	tacagcttctggccttgttg	59,07	yes
BPV1	5856	6010	Random Frag 6	M13-caggactgttcacaacccaag	60,59	gggtggtgacttttctattcaca	60,27	none
BPV1	6519	6693	C+1 6589	M13-ttaatcggccctactggcta	59,7	tggtatacattgaattttgagctatca	60,15	none
BPV1	6638	6790	Random Frag 3	M13-cctcagatggaaccccacta	59,92	agggcataagtccttgcaga	59,84	none
BPV1	6968	7154	LCR P1	M13-tggacttagatcaatttcccttg	59,49	aggtgcagttgacttaccttctg	59,86	none
BPV1	7061	7263	LCR P2	M13-cttccagtaagcctgcaaaaa	59,52	gccagcacaataattcaatgc	60,48	none
BPV1	7205	7354	Random Frag 1	M13-acacccggtacacatcctgt	60,16	gcgtcccattgatgcttagt	60,1	none
BPV1	7205	7391	LCR P3	M13-acacccggtacacatcctgt	60,16	tacttaccccgcatctaccg	59,97	none
BPV1	7325	7520	LCR P4	M13-ttggcaagaaactaagcatcaa	59,89	accagttctggttccgacag	60,15	none
BPV1	7486	7656	LCR P5	M13-tctatttttgtctctctgtcggaac	60,17	gggagccccaaacctatatc	59,63	none
BPV1	7598	7794	LCR P6	M13-atggtgcatagcggatgtct	60,51	agcaccgaagacggtttg	59,83	weak
BPV1	7734	7904	LCR P7	M13-aagttgtaacctgatctccacaaag	59,98	cggtttcggtgagcttaaaa	60,24	none
BPV1	7857	81	LCR P8	M13-cgggagccaatcaaaatg	60	ggggtctgtcagcagcttt	60,4	none

Table 41) Primers for EMSA testing of BPV1 part II

Template	Start	End	Fragment name	Sequence FW 5' to 3' with M13 (-21) tag at 5'	Tm FW	Sequence RV 5' to 3'	Tm RV	CTCF binding
HPV18	754	943	HPV18 pred.844	M13-accacaacgtcacacaatgtt	58,82	ctccccgtctgtaccttctg	59,72	none
HPV18	1102	1297	HPV18 E+1 1205	M13-agagacagcacaggcattgtt	59,93	ccgccttttttgccttttt	60,18	none
HPV18	2926	3117	HPV18 cor. C+1 2990	M13-ggcaactaatacgttgggaaaa	60,23	tgtcttgcagtgtccaatcc	59,68	strong
HPV18	3381	3575	HPV18 E-2 3487	M13-tgggaagtacattttgggaataa	59,61	tccacagtgtccaggtcgt	60,15	none
HPV18	3527	3718	HPV18 pred.3621	M13-aaagacctacggccagacg	60.26	cattttaaactgtttctgtcacctttt	59.15	medium
HPV18	4440	4638	HPV18 pred.4505	M13-ggggtcgtacagggtacatt	58.65	gatgttatatcaaacccagacgtg	59.67	two medium bands
HPV18	4947	5155	HPV18 CTCF?A 4947-5155	M13-cagtggctaaccctgagtttct	59,81	agttgcccgttgacctaatc	59,06	none
HPV18	5045	5253	HPV18 CTCF?B 5045-5253	M13-cgtagtgatgttcctgattcagat	59,55	ggctgcagttcaatatattctgg	60	none
HPV18	5381	5577	HPV18 E-1 5475	M13-tctgcctcttcctatagtaatgtaacg	60,41	ggaataaaataatataatggccacaaa	59,7	medium
HPV18	5655	5850	HPV18 pred.5768	M13-cctccttctgtggcaagagt	59.45	ggtcaggtaactgcaccctaa	59.11	medium

Table 42) Primers for EMSA testing of HPV18

Template	Start	End	Fragment name	Sequence FW 5' to 3' with M13 (-21) tag at 5'	Tm FW	Sequence RV 5' to 3'	Tm RV	CTCF binding
HPV 16	1212	1405	HPV16 C+2 1283	M13-aaacaaagtagagctgcaaaaagg	60,32	ctaacaccctctcccccact	60,36	none
HPV 16	2852	3049	HPV16 C+1 2916	M13-ggaaacacatgcgcctaga	59,81	ggctaacgtcttgtaatgtccac	59,95	strong
HPV 16	4999	5206	HPV16 E-1 5119	M13-ctgcatatgaaggtatagatgtggat	59,71	accttagcacctatagattttccact	59,15	weak
HPV 16 L1 plasmid	6051	6278	HPV16 E+1 6127	M13-tgcaggtgtggataatagagaatg	60,38	tgtaatgtagtaaagtccatagcacca	60,23	weak
HPV 16	6426	6600	HPV16 C-1 6512	M13-ggctggtactgttggtgaaaa	60,02	attattgtggccctgtgctc	59,96	weak
HPV 16	6772	6957	HPV16 pred.6860	M13-tgcaaaataaccttaactgcagac	59.74	gggatcatcttctttaggtgct	58.74	weak

Table 43) Primers for EMSA testing of HPV16

The HPV16 fragment HPV16 E+1 6127 needed to be amplified from a separate template plasmid since it includes the site were the other HPV16 template plasmid was cut to be inserted into an amplification vector.

Template	Start	End	Fragment name	Sequence FW 5' to 3' with M13 (-21) tag at 5'	Tm FW	Sequence RV 5' to 3'	Tm RV	CTCF binding
HPV16 114/K	1216	1406	HPV16 114K pred.1284	M13-caaagtagagctgcaaaaagga	60,32	ctaacaccctctcccccact	60,36	none
HPV16 114/K	2853	3050	HPV16 114K C+1 2917	M13-ggaaacacatgcgcctaga	59,81	ggctaacgtcttgtaatgtccac	59,95	strong
HPV16 114/K	5000	5207	HPV16 114K E-1 5120	M13-ctgcatatgaaggtatagatgtggat	59,71	accttagcacctatagattttccact	59,15	weak
HPV 16 L1 plasmid	6052	6279	HPV16 114K E+1 6128	M13-tgcaggtgtggataatagagaatg	60,38	tgtaatgtagtaaagtccatagcacca	60,23	weak
HPV16 114/K	6421	6621	HPV16 114K pred.6516	M13-taatagggctggtgctgttg	58.80	taacaccctctcccccactt	60.74	weak
HPV16 114/K	6773	6958	HPV16 114K pred.6861	M13-tgcaaaataaccttaactgcagac	59.74	gggatcatcttctttaggtgct	58.74	weak

Table 44) Primers for EMSA testing of HPV16 114/K

The HPV16 114/K fragment HPV16 114/K E+1 6127 needed to be amplified from a separate template plasmid since it includes the site were the other HPV16 114/K template plasmid was cut to be inserted into an amplification vector. The predicted binding on this fragment is identical with the one in HPV16. Thus the HPV16 L1 plasmid could be used as template.

Template	Start	End	Fragment name	Sequence FW 5' to 3' with M13 (-21) tag at 5'	Tm FW	Sequence RV 5' to 3'	Tm RV	CTCF binding
HPV31	534	713	HPV31 pred.616	M13-cctcgtactgaaacccaagtg	59.64	aattggatgtgtccggttct	59.26	none
HPV31	804	1008	HPV31 E+4 885	M13-tgttaatgggctcatttggaa	60,31	gtcaaccatatcctccccagt	60,07	none
HPV31	1029	1200	HPV31 E+2 1093	M13-atacaacaatcaggcagaagca	59,77	gcatatagcttttaaccgtggact	59,95	none
HPV31	1182	1374	HPV31 C+2/E+3 1278	M13-acggttaaaagctatatgcatagaaaa	59,89	tgttggagtttcattctctcgtt	60,16	none
HPV31	2230	2406	HPV31 second C+3 2332	M13-ggtgcacctaatacaggtaaatca	59,35	gtaattgtctatataatgccaacatgg	59,12	none
HPV31	2357	2531	HPV31 second C+1 2413	M13-gcatgttagatgatgctacaacg	59,69	ggccatctgtcatccttacct	60,34	strong
HPV31	2801	3015	HPV31 A? 2801-3015	M13-cgacttgaatgtgtattaatgtataaagc	59,73	cctgtaggtgcagttaaatacagttc	59,59	strong
HPV31	2894	3093	HPV31 B? 2894-3093	M13-aaagccttacaagctattgaactaca	59,08	cagttagtataatgcatggtgttgtg	59,79	none
HPV31	5077	5273	HPV31 E+1 5179	M13-caaactttgcgcactcgtag	59,67	gtggcaggtgtatccacagtaa	59,92	strong
HPV31	6354	6540	HPV31 pred.6432	M13-ttggtgaatcggtccctact	59.41	aactgattgccccaacaaat	59.29	medium

Table 45) Primers for EMSA testing of HPV31

Template	Start	End	Fragment name	Sequence FW 5' to 3' with M13 (-21) tag at 5'	Tm FW	Sequence RV 5' to 3'	Tm RV	CTCF binding
HPV6b	1251	1460	HPV6b E+2 1357	M13-tgaagtggaagctggaacg	59,97	ccaagtaatgctgcccgta	59,7	none
HPV6b	2801	3007	HPV 6b A? 2801-3007	M13-cacaaacatgtattgcattgga	59,35	ttgtaatgtccacggttccat	60,1	none
HPV6b	2887	3101	HPV 6b B? 2887-3101	M13-aatgcaagtagtgccaccatta	59,54	tgtttgcacagccatcaaat	60,12	none
HPV6b	4715	4913	HPV6b E-5 4790	M13-ccccttcggatccatctatt	60,11	gaaggttctgtaaagacaggatttct	59,64	none
HPV6b	4913	5102	HPV6b C-2 5018 E+1 4987	M13-ctgtaacacaaccccaaccac	60,18	cgactatataggcccacacga	59,99	none
HPV6b	5317	5515	HPV6b C-1 5425 E-1 5431	M13-attggacaacgggggtctat	60,44	taacagggtgttgggtaggg	59,71	strong
HPV6b	5995	6199	HPV6b E+3 6099	M13-atttaaggtggtgttaccagatcc	59,57	ctgtccagggttaccaccac	60,28	none
HPV6b	6179	6382	HPV6b E-2 6264	M13-agtggtggtaaccctggaca	60,28	gtcaaccatatcgccatcct	59,78	medium
HPV6b	7155	7380	HPV6b E-3 7206 E-4 7257	M13-agtatcctttgggacgcaag	59,19	caaataacttacattacacacagaacaca	59,78	none

Table 46) Primers for EMSA testing of HPV6b

Template	Start	End	Fragment name	Sequence FW 5' to 3' with M13 (-21) tag at 5'	Tm FW	Sequence RV 5' to 3'	Tm RV	CTCF binding
HPV11	1295	1494	HPV11 E+1 1357	M13-aaaatgggggagatggtca	60,12	aacaaatgacagcccaaagc	60,12	none
HPV11	2801	3003	HPV 11 A? 2801-3003	M13-cacaaacacattatgcattgga	59,35	aatgtccaaggttccacacc	59,68	none
HPV11	2900	3104	HPV 11 B? 2900-3104	M13-ccaccattaactgtgtcagagact	59,61	cattgtcttcacagccatcaa	59,71	none
HPV11	3930	4158	HPV11 E-3 4058	M13-tgttattgcatttgcagtatgtttt	59,86	atcaccattattgttgcgtttg	59,76	medium
HPV11	4709	4898	HPV11 E-2 4781	M13-cttccgatccctccattgt	59,87	tctgtaaacagtggattttgaaaca	59,96	none
HPV11	4844	5041	HPV11 pred.4921	M13-tgtctgttaccaatcacactacca	59.47	tggatgtaggtccactatcactaga	59.10	weak
HPV11	5330	5501	HPV11 C-1 5416 E-1 5422	M13-acacacgcagtggacaacat	60,08	gaatgttggacagggtcagg	60,36	strong
HPV11	6243	6428	HPV11 C-2 6311	M13-gtgctccaccgttaggtga	59,69	caaggggaacatccgattta	59,76	weak
HPV11	6544	6738	HPV11 pred.6636	M13-taatagggccggtactgtgg	59.84	ccccagcaaataccattgtt	59.69	weak, smear
HPV11	6872	7074	HPV11 pred.6980	M13-ttacagtttatttttcaattgtgtagca	59.52	tcctgtttttctttttcaggtgt	59.21	none

Table 47) Primers for EMSA testing of HPV11

Template	Start	End	Fragment name	Sequence FW 5' to 3' with M13 (-21) tag at 5'	Tm FW	Sequence RV 5' to 3'	Tm RV	CTCF binding
HPV38	35	219	HPV38 C-1 117	M13-agacttttctctttaaccgtaggc	59,77	tgaggttttggtagttccatga	59,47	medium
HPV38	176	363	HPV38 E+1 254	M13-ggcctgtaagcttgggatg	60,61	tcctcttgagtccagattaactgtaag	60,53	none
HPV38	584	791	HPV38 pred.690	M13-aattggaaaggaaggtgcag	59.17	cccccacaaagaactatgattt	59.24	none
HPV38	1243	1431	HPV38 C+3 1331	M13-tgtggagcaagacagtggac	59,87	ccaaacgagtctttaaatttgctt	60,06	medium
HPV38	3436	3666	HPV38 C+1/2 and E+2/4/5 3540	M13-aaaccacaggcacccaaa	59,94	ccactttgtcaggcgatatg	59,15	none
HPV38	4357	4545	HPV38 E+3 4454	M13-accaaccgaactgatacctattg	59,32	gtattgcggcctccactc	59,2	none

Table 48) Primers for EMSA testing of HPV38

Template	Start	End	Fragment name	Sequence FW 5' to 3' with M13 (-21) tag at 5'	Tm FW	Sequence RV 5' to 3'	Tm RV	CTCF binding
BPV1	4910	5051	Neg. control	M13-CCCGCAGTATTGCCTCTAAA	60,22	AGCACAGCTGGTTCTGCTTC	60,74	none
Human C-myc promoter	41894249	41894403	Pos. control	M13-GGGATCGCGCTGAGTATAAA	58,74	GGATCTCCCTTCCCAGGAC	59,24	strong
All forward primers	41894249	41894403	FAM-M13(-21)	FAM-TGTAAAACGACGGCCAGT	57,7	NA	NA	NA
M13 tag	NA	NA	M13 sequence preceding every FW primer	TGTAAAACGACGGCCAGT	NA	NA	NA	NA

Table 49) Primers for EMSA used to generate control fragments or label all fragments with FAM

6.4 Electrophoretic mobility shift assays of all fragments tested

6.4.1 BPV1

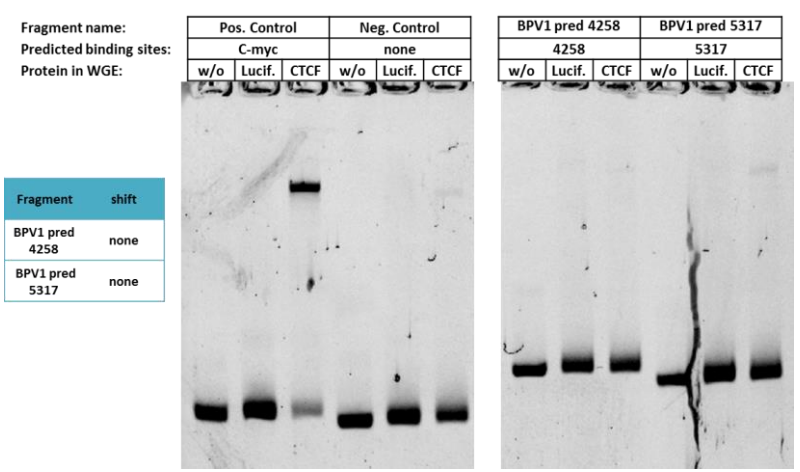
	Fragment Start	Fragment End	Predicted motifs on fragment	Tool	Fragment name	CTCF band shift
Fragments of *Storm* predictions	4145	4354	4258	4PWM [7.1]	BPV1 pred.4258	none
	5232	5422	5317	6PWM [10.6]	BPV1 pred.5317	none
Fragments of CTCFBSDB 1.0 and Essex tool predictions	6519	6693	6589	CTCFBSDB [5.9]	C+1 6589	none
	3576	3775	C 3671, S 3670	5PWM [10.3], 6PWM [8.5],CTCFBSDB [10,3]	C-1 3671	weak
	2917	3096	3025	Essex	Essex+1 3025	strong
	2327	2488	2427, 2397	Essex	Essex+2 2427, Essex-2 2397	weak
	4442	4611	4534	Essex	Essex+3 4534	none
	580	745	663	Essex	Essex-1 663	none
	1137	1309	1227	Essex	Essex-3 1227	two bands, weak
	5105	5263	5190	Essex	Essex-4 5190	none
	4689	4863	4775	Essex	Essex-5 4775	medium
	4049	4226	4144	Essex	Essex-6 4144	none
Fragments covering functional elements	6968	7154	none	N/A	LCR P1	none
	7061	7263	none	N/A	LCR P2	none
	7205	7391	none	N/A	LCR P3	none
	7325	7520	none	N/A	LCR P4	none
	7486	7656	none	N/A	LCR P5	none
	7598	7794	none	N/A	LCR P6	weak
	7734	7904	none	N/A	LCR P7	none
	7857	81	none	N/A	LCR P8	none
	55	223	none	N/A	LCR P9	none
	772	950	none	N/A	P890 region	weak
	1031	1201	none	N/A	E2 binding site at 1125	none

Table 50) Part I of the BPV1 fragments of the EMSAs images below

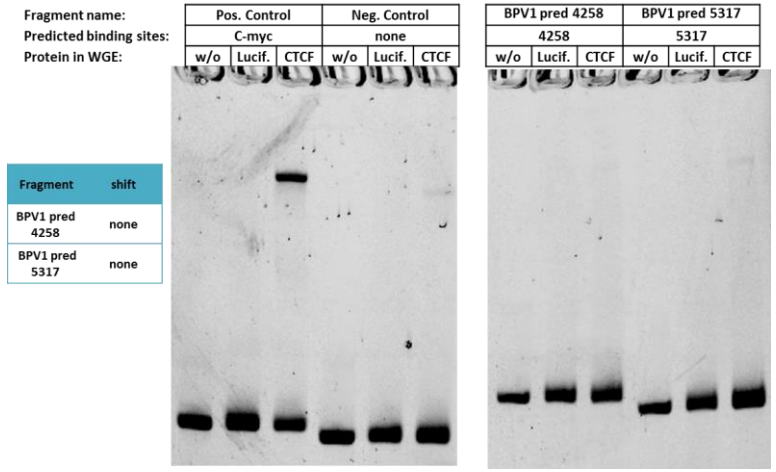
	Fragment		Predicted motifs on fragment	Tool	Fragment name	CTCF band shift
	Start	End				
Random fragments	7205	7354	none	N/A	Random Frag 1	none
	5475	5639	none	N/A	Random Frag 2	weak
	6638	6790	none	N/A	Random Frag 3	none
	2650	2819	none	N/A	Random Frag 4	none
	3072	3238	none	N/A	Random Frag 5	none
	5856	6010	none	N/A	Random Frag 6	none
	843	983	none	N/A	Random Frag 7	none
Overlapping fragments at 3025	2704	2915	none	N/A	No CTCF	none
	2785	2984	none	N/A	Frag CTCF?A	none
	2831	3027	none	N/A	Frag CTCF?B	strong
	2917	3096	3025	Essex	E+1 3025 (at 5’ edge)	strong
	2947	3120	3025	Essex	E+1 in middle	none
	2995	3174	3025	Essex	E+1 at 3’ edge	none

Table 51) Part II of the BPV1 fragments of the EMSAs images below

6.4.1.1 Fragments of *Storm* predictions

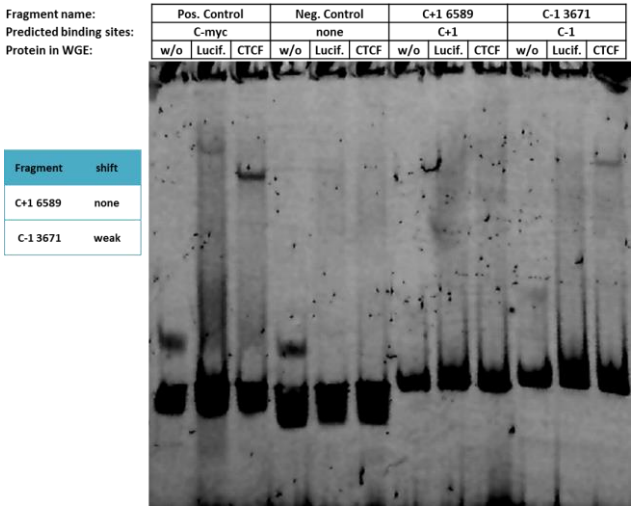


Fragment	shift
BPV1 pred 4258	none
BPV1 pred 5317	none

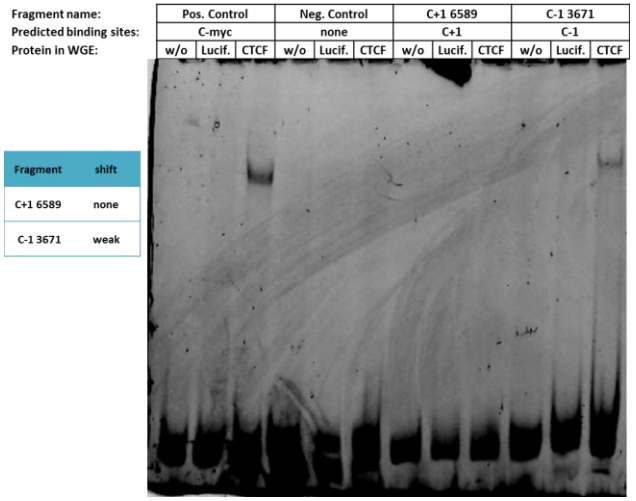


Fragment	shift
BPV1 pred 4258	none
BPV1 pred 5317	none

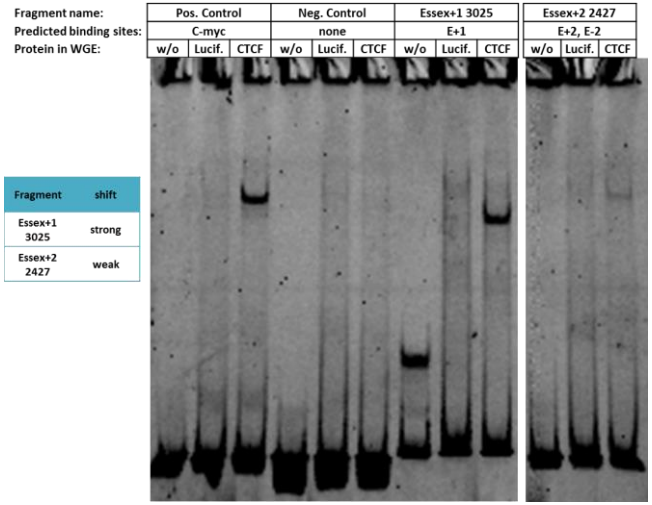
6.4.1.2 Fragments of CTCFBSDB 1.0 and Essex tool predictions



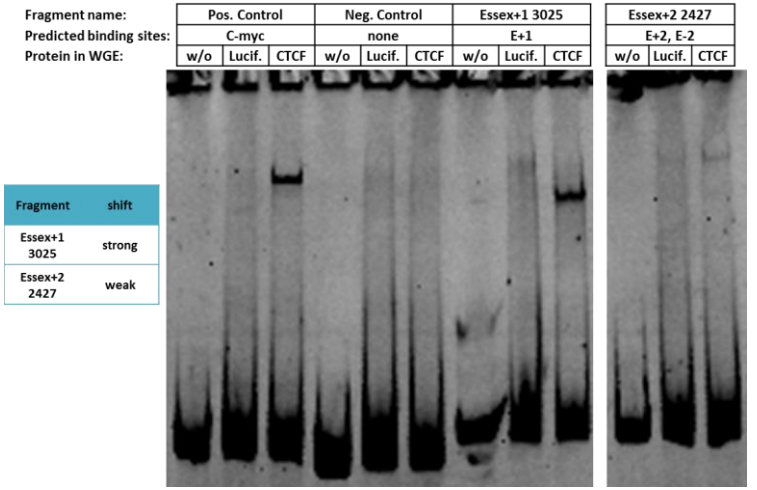
Fragment	shift
C+1 6589	none
C-1 3671	weak



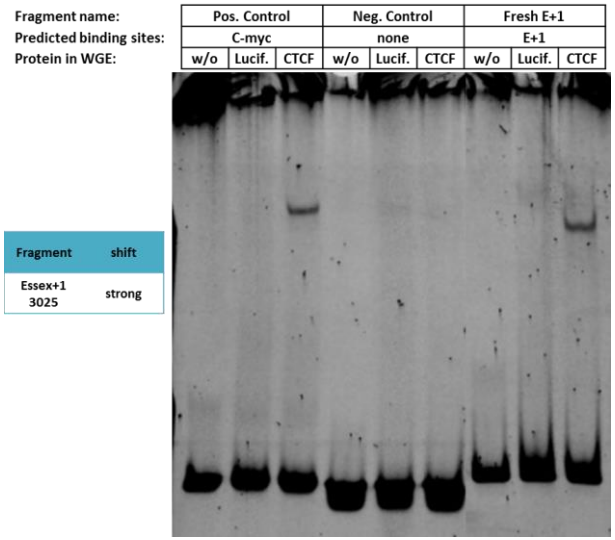
Fragment	shift
C+1 6589	none
C-1 3671	weak

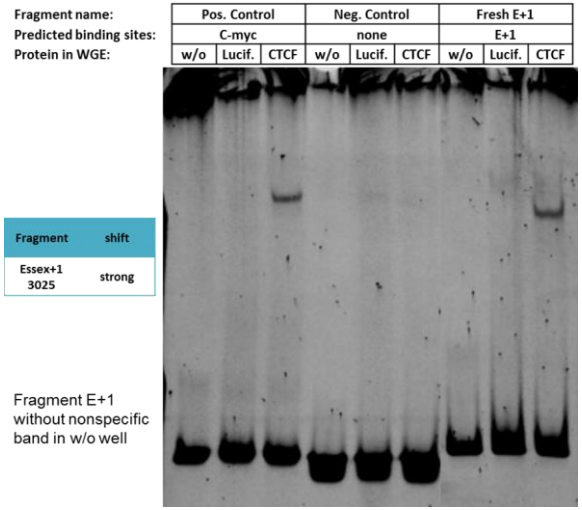


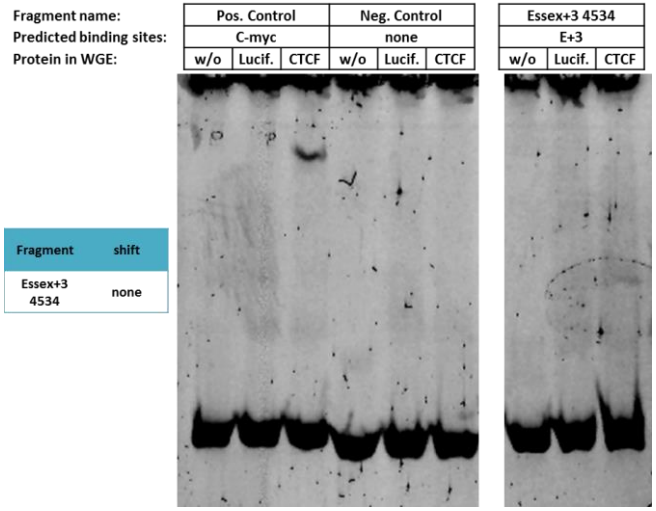
Fragment	shift
Essex+1 3025	strong
Essex+2 2427	weak

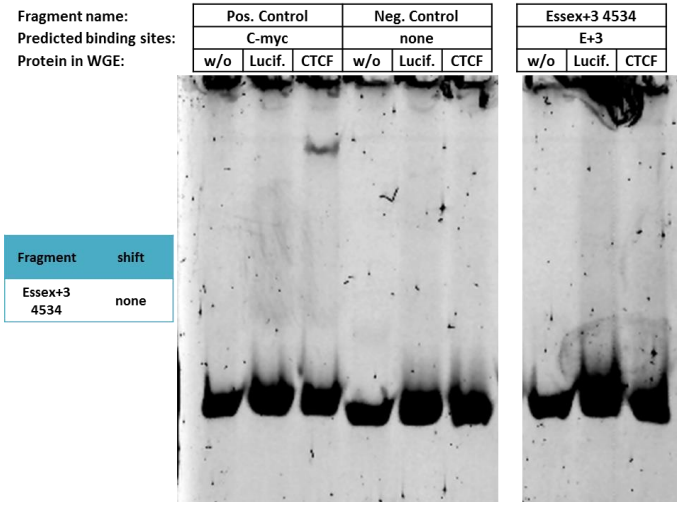


Fragment	shift
Essex+1 3025	strong
Essex+2 2427	weak

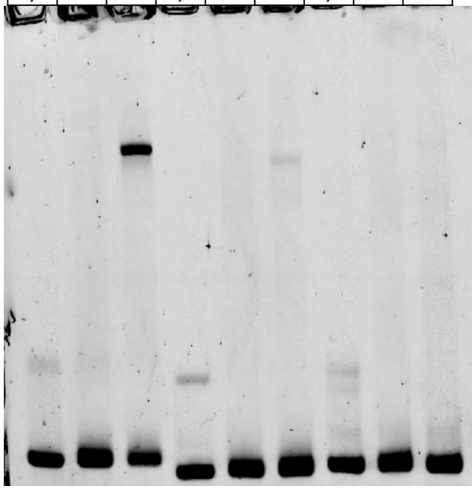

Fragment name:
Predicted binding sites:
Protein in WGE:
Pos. Control
C-myc
w/o
Lucif.
CTCF
Neg. Control
none
Fresh E+1
E+1
Fragment
shift
Essex+1 3025
strong


Fragment name:
Predicted binding sites:
Protein in WGE:
Pos. Control
C-myc
w/o
Lucif.
CTCF
Neg. Control
none
Fresh E+1
E+1
Fragment
shift
Essex+1 3025
strong
Fragment E+1 without nonspecific band in w/o well


Fragment name:
Predicted binding sites:
Protein in WGE:
Pos. Control
C-myc
w/o
Lucif.
CTCF
Neg. Control
none
Essex+3 4534
E+3
Fragment
shift
Essex+3 4534
none

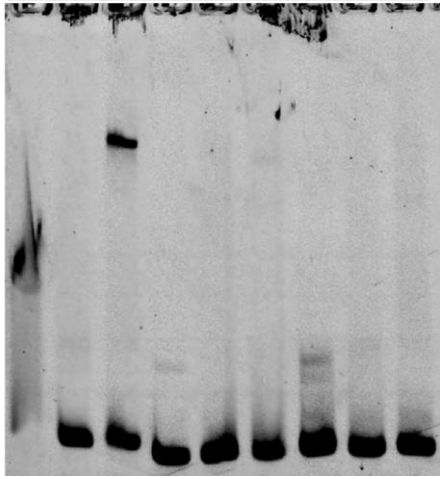

Fragment name:
Predicted binding sites:
Protein in WGE:
Pos. Control
C-myc
w/o
Lucif.
CTCF
Neg. Control
none
Essex+3 4534
E+3
Fragment
shift
Essex+3 4534
none

Fragment name:	Pos. Control			Neg. Control			Essex-1 663		
Predicted binding sites:	C-myc			none			E-1		
Protein in WGE:	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF



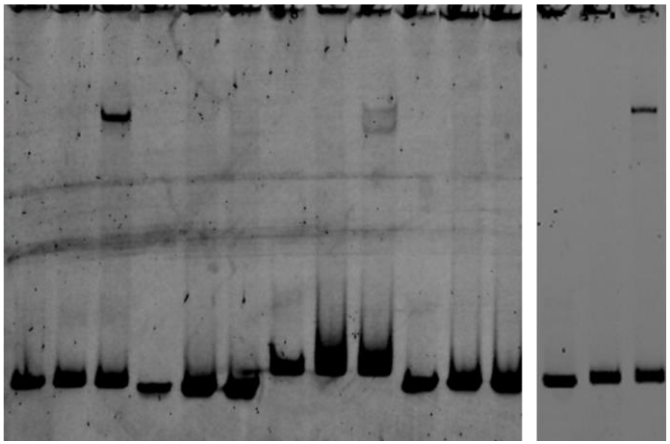
Fragment	shift
Essex-1 663	none

Fragment name:	Pos. Control			Neg. Control			Essex-1 663		
Predicted binding sites:	C-myc			none			E-1		
Protein in WGE:	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF



Fragment	shift
Essex-1 663	none

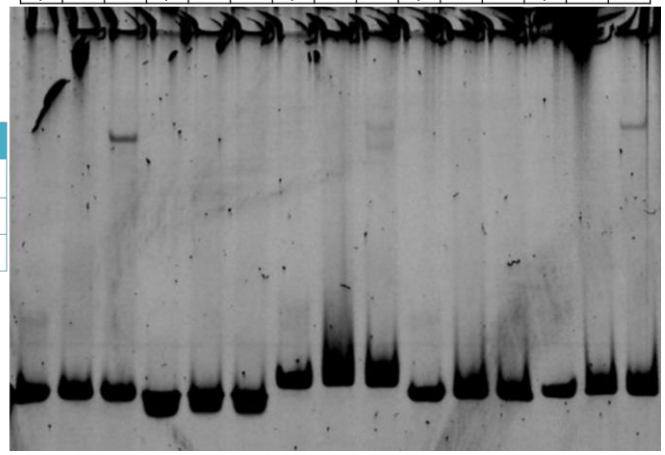
Fragment name:	Pos. Control			Neg. Control			Essex-3 1227			Essex-4 5190			Essex-5 4775		
Predic. binding sites:	C-myc			none			E-3			E-4			E-5		
Protein in WGE:	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF



Fragment	shift
Essex-3 1227	Two weak bands
Essex-4 5190	none
Essex-5 4775	medium

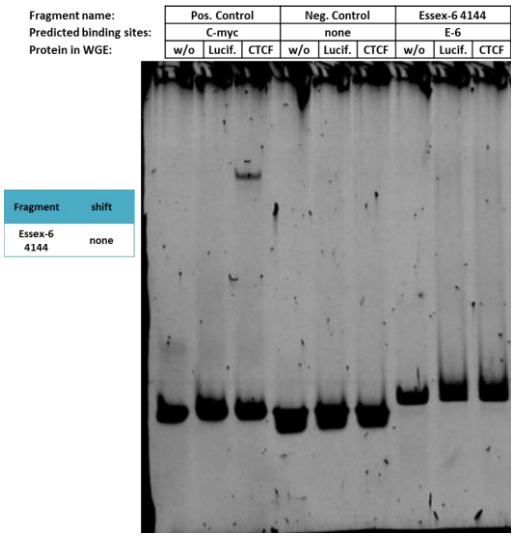
Comments:
E-2 tested together with E+2 on earlier slide

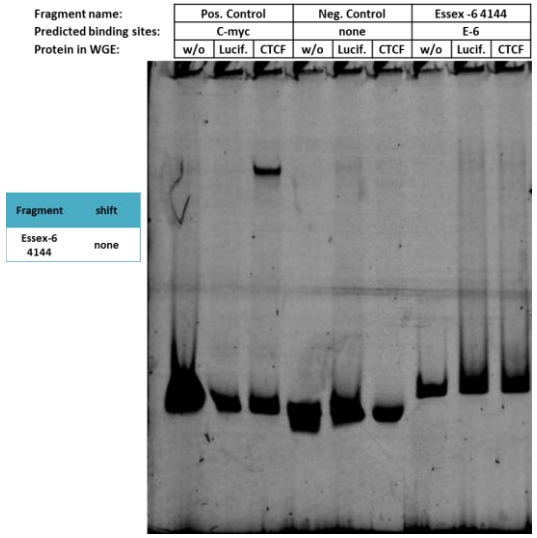
Fragment name:	Pos. Control			Neg. Control			Essex-3 1227			Essex-4 5190			Essex-5 4775		
Predicted binding sites:	C-myc			none			E-3			E-4			E-5		
Protein in WGE:	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF



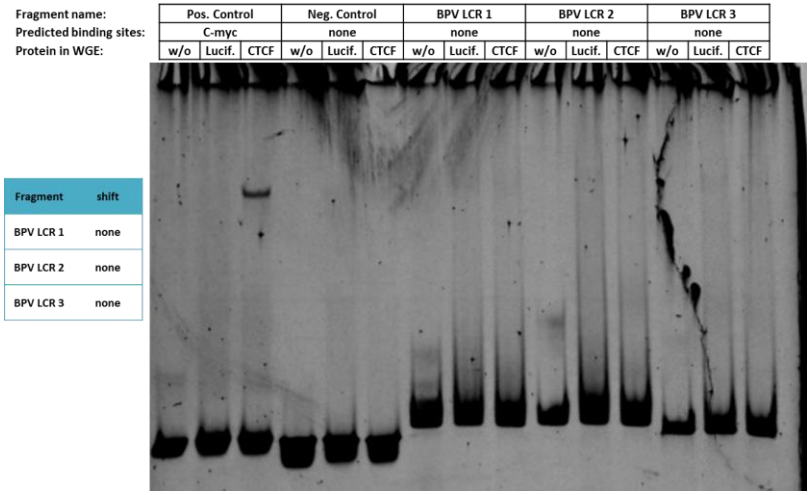
Fragment	shift
Essex-3 1227	Two weak bands
Essex-4 5190	none
Essex-5 4775	medium

Comments:
E-2 tested together with E+2 on earlier slide

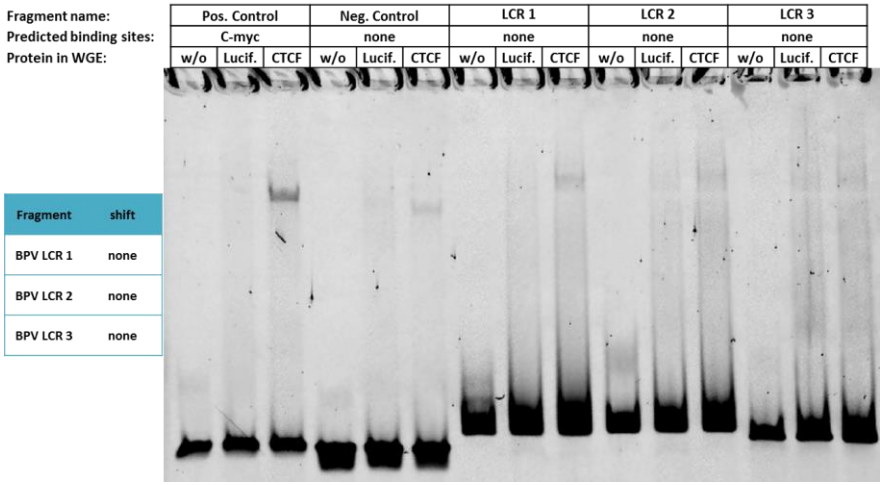

Fragment name:
Pos. Control
Neg. Control
Essex-6 4144
Predicted binding sites:
C-myc
none
E-6
Protein in WGE:
w/o
Lucif.
CTCF
w/o
Lucif.
CTCF
w/o
Lucif.
CTCF
Fragment
shift
Essex-6 4144
none


Fragment name:
Pos. Control
Neg. Control
Essex -6 4144
Predicted binding sites:
C-myc
none
E-6
Protein in WGE:
w/o
Lucif.
CTCF
w/o
Lucif.
CTCF
w/o
Lucif.
CTCF
Fragment
shift
Essex-6 4144
none

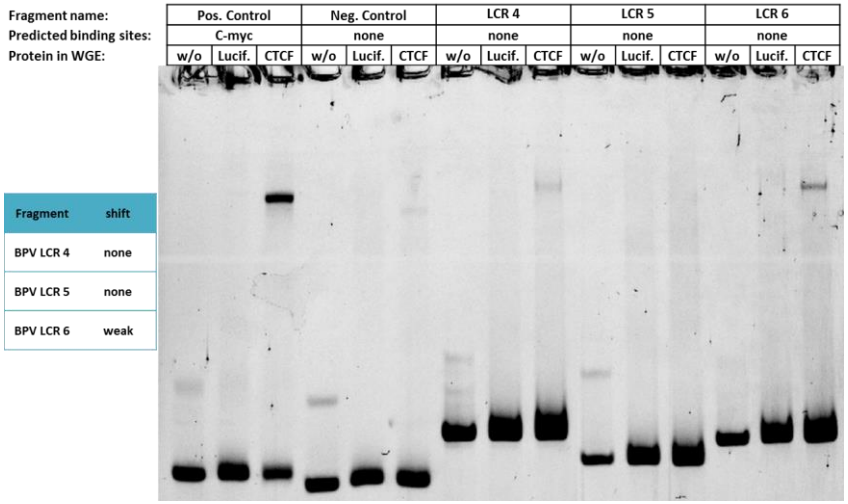
6.4.1.3 Fragments covering functional elements



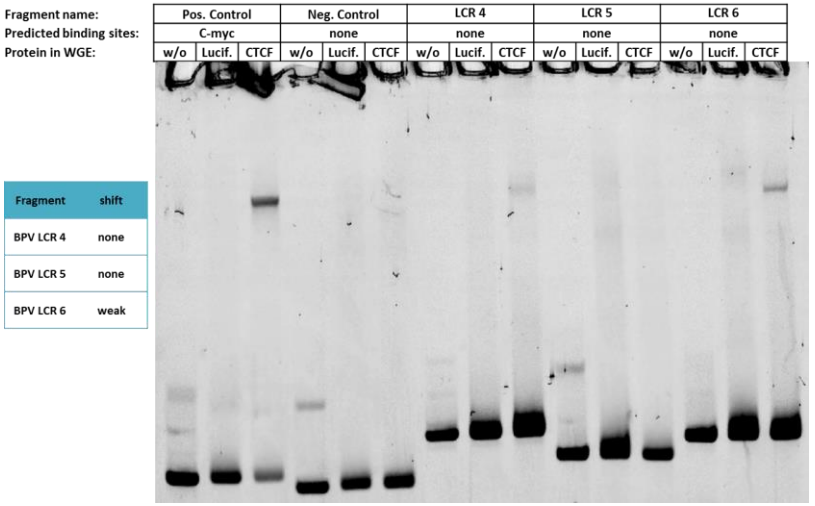
Fragment	shift
BPV LCR 1	none
BPV LCR 2	none
BPV LCR 3	none



Fragment	shift
BPV LCR 1	none
BPV LCR 2	none
BPV LCR 3	none

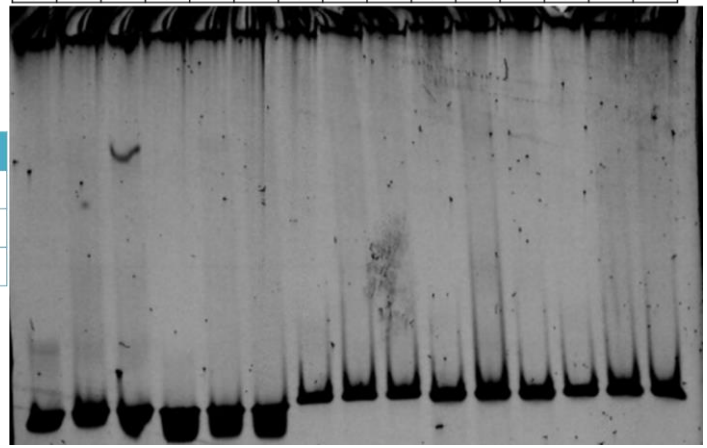


Fragment	shift
BPV LCR 4	none
BPV LCR 5	none
BPV LCR 6	weak



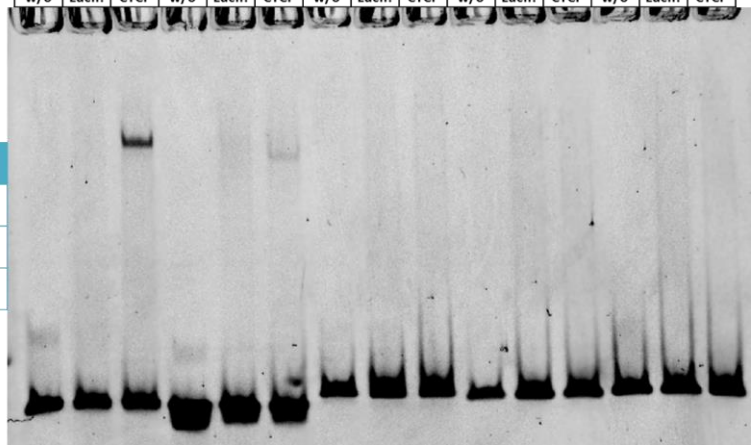
Fragment	shift
BPV LCR 4	none
BPV LCR 5	none
BPV LCR 6	weak

Fragment name:	Pos. Control			Neg. Control			BPV LCR 7			BPV LCR 8			BPV LCR 9		
Predicted binding sites:	C-myc			none			none			none			none		
Protein in WGE:	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF



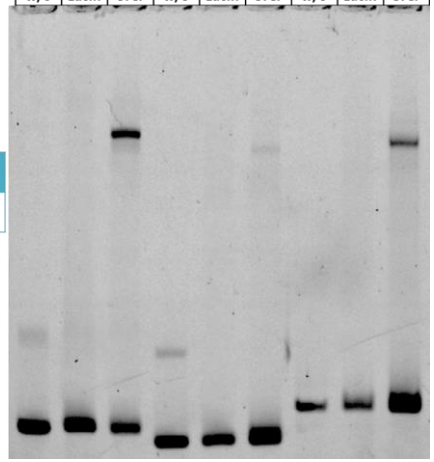
Fragment	shift
BPV LCR 7	none
BPV LCR 8	none
BPV LCR 9	none

Fragment name:	Pos. Control			Neg. Control			LCR 7			LCR 8			LCR 9		
Predicted binding sites:	C-myc			none			none			none			none		
Protein in WGE:	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF



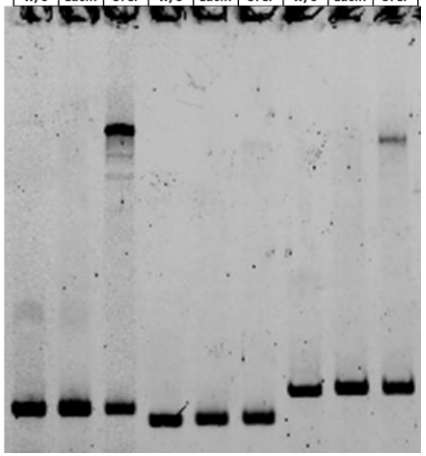
Fragment	shift
BPV LCR 7	none
BPV LCR 8	none
BPV LCR 9	none

Fragment name:	Pos. Control			Neg. Control			P890		
Predicted binding sites:	C-myc			none			none		
Protein in WGE:	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF

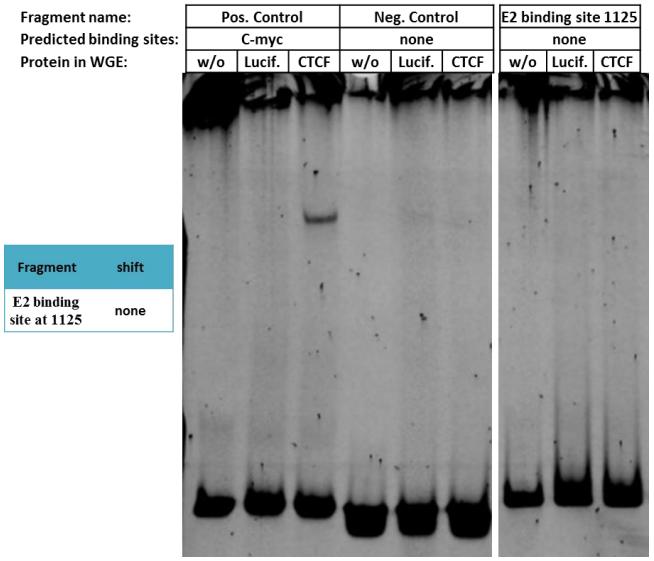


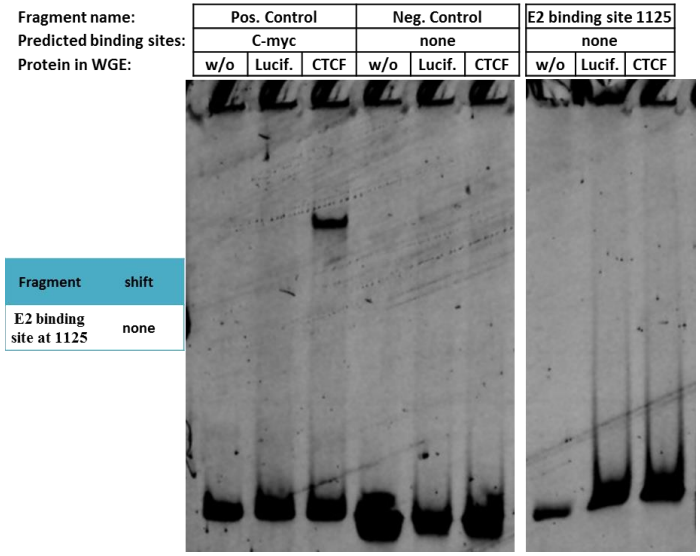
Fragment	shift
P890	medium

Fragment name:	Pos. Control			Neg. Control			P890		
Predicted binding sites:	C-myc			none			none		
Protein in WGE:	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF

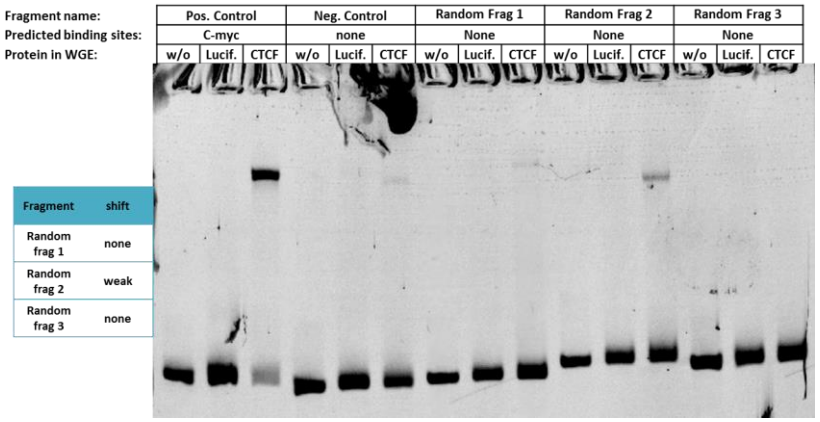


Fragment	shift
P890	medium

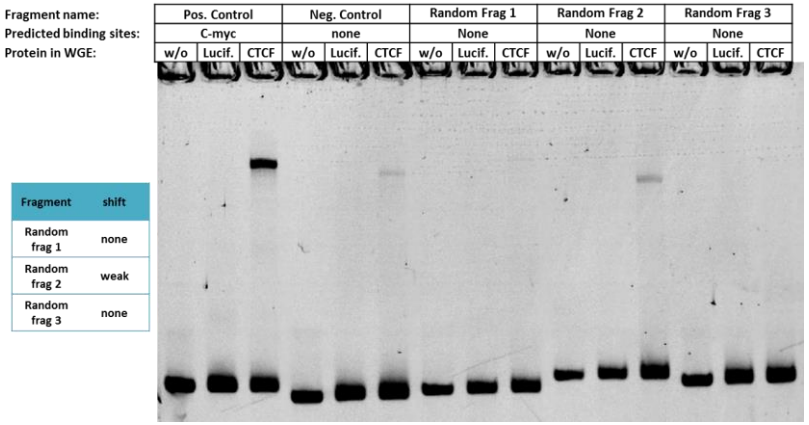

Fragment name:
Predicted binding sites:
Protein in WGE:
Pos. Control
C-myc
w/o
Lucif.
CTCF
Neg. Control
none
w/o
Lucif.
CTCF
E2 binding site 1125
none
w/o
Lucif.
CTCF
Fragment
shift
E2 binding site at 1125
none


Fragment name:
Predicted binding sites:
Protein in WGE:
Pos. Control
C-myc
w/o
Lucif.
CTCF
Neg. Control
none
w/o
Lucif.
CTCF
E2 binding site 1125
none
w/o
Lucif.
CTCF
Fragment
shift
E2 binding site at 1125
none

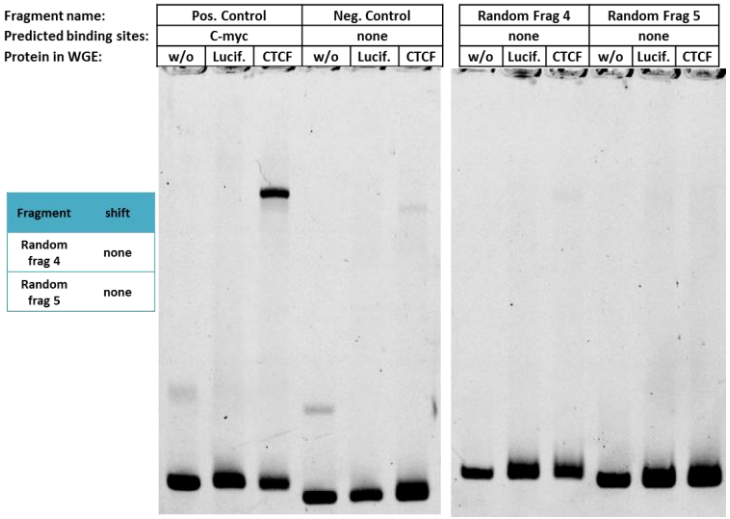
6.4.1.4 Random fragments



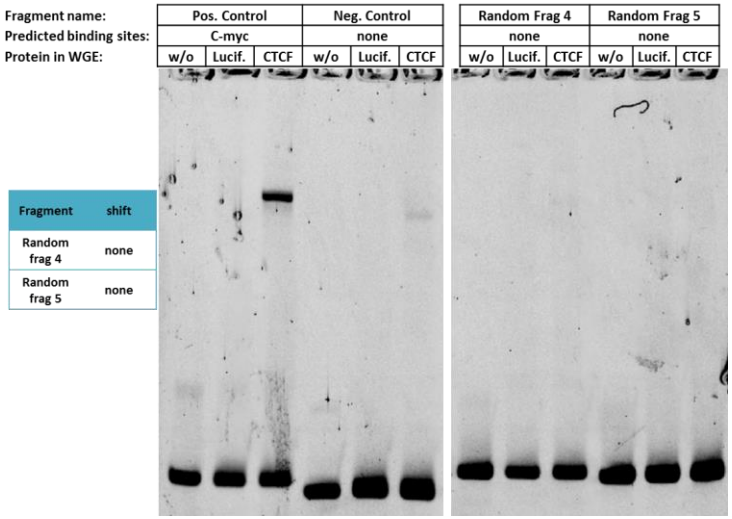
Fragment	shift
Random frag 1	none
Random frag 2	weak
Random frag 3	none



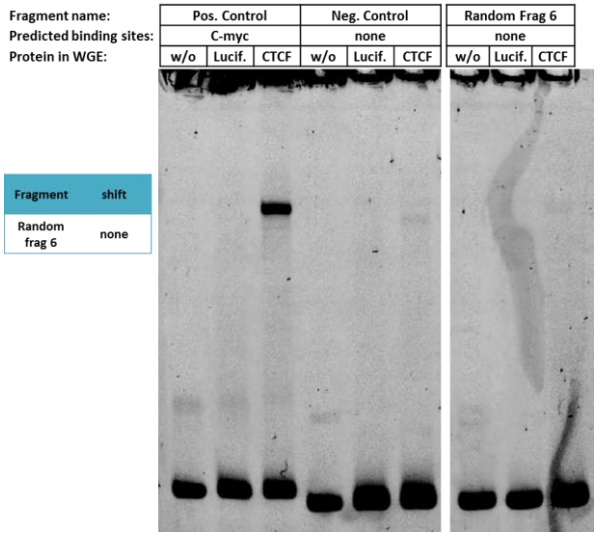
Fragment	shift
Random frag 1	none
Random frag 2	weak
Random frag 3	none

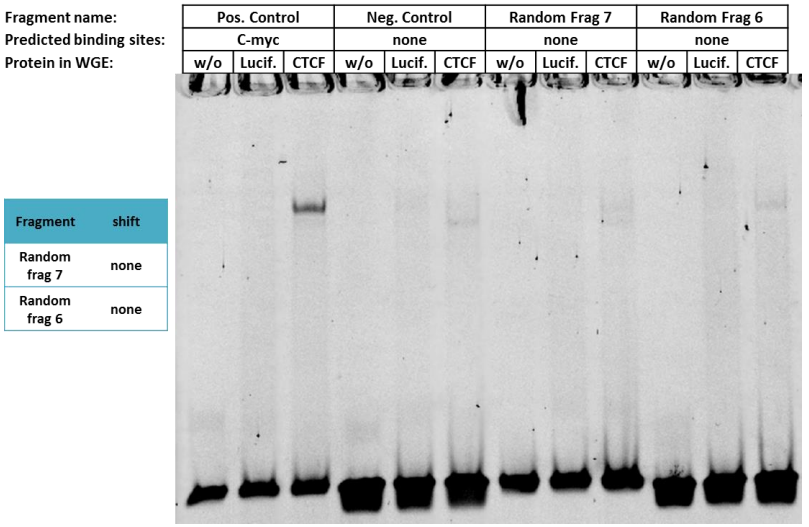


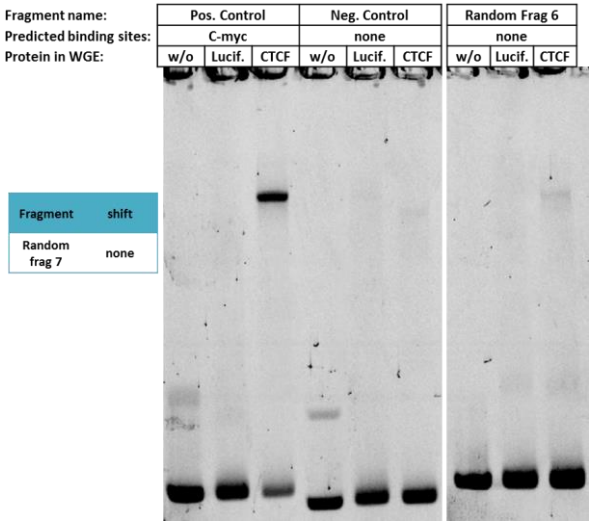
Fragment	shift
Random frag 4	none
Random frag 5	none



Fragment	shift
Random frag 4	none
Random frag 5	none


Fragment name:
Predicted binding sites:
Protein in WGE:
Pos. Control
C-myc
w/o
Lucif.
CTCF
Neg. Control
none
Random Frag 6
Fragment
shift
Random frag 6
none


Fragment name:
Predicted binding sites:
Protein in WGE:
Pos. Control
C-myc
Neg. Control
Random Frag 7
Random Frag 6
none
w/o
Lucif.
CTCF
Fragment
shift
Random frag 7
none
Random frag 6
none

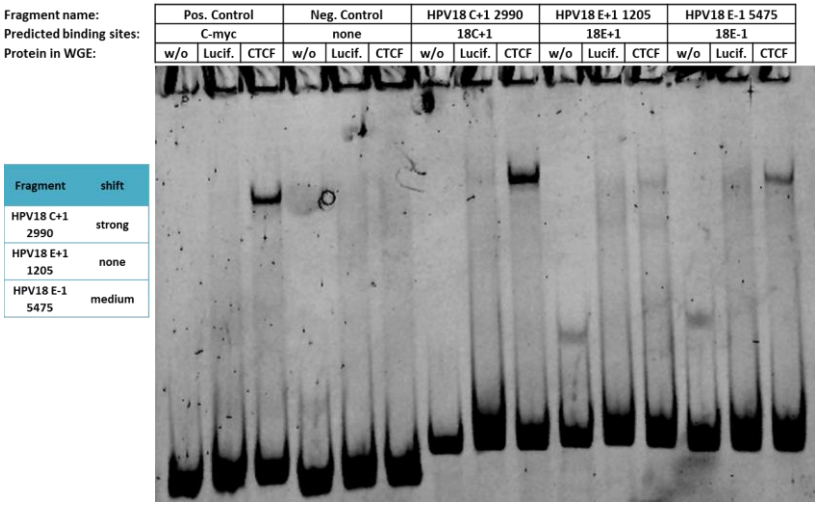

Fragment name:
Predicted binding sites:
Protein in WGE:
Pos. Control
C-myc
w/o
Lucif.
CTCF
Neg. Control
none
Random Frag 6
Fragment
shift
Random frag 7
none

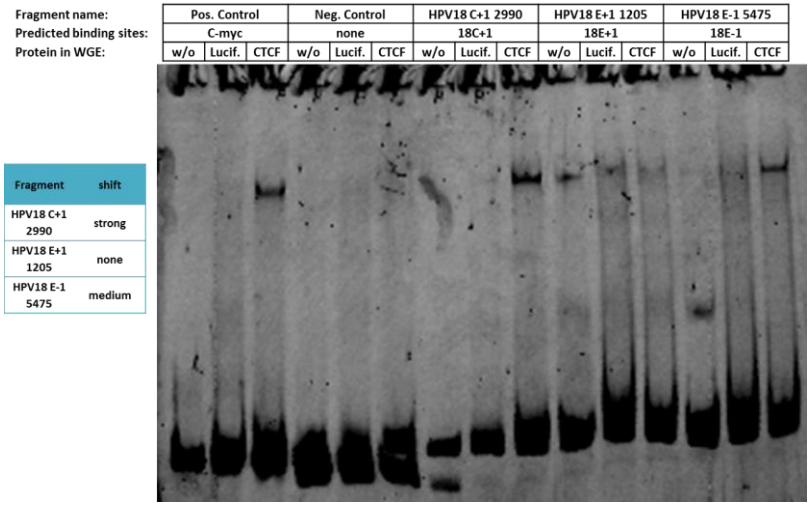
6.4.2 High risk HPV types

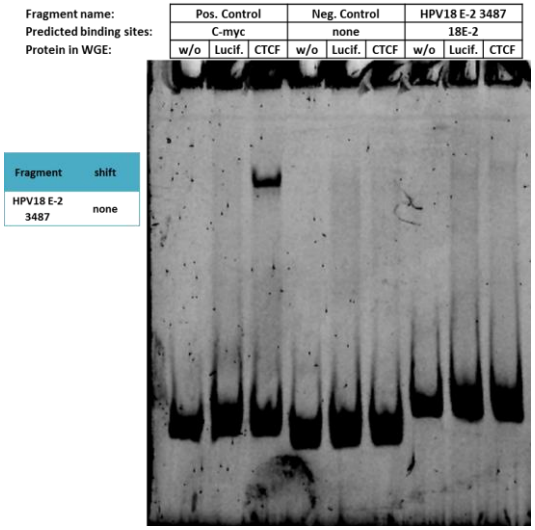
6.4.2.1 HPV18

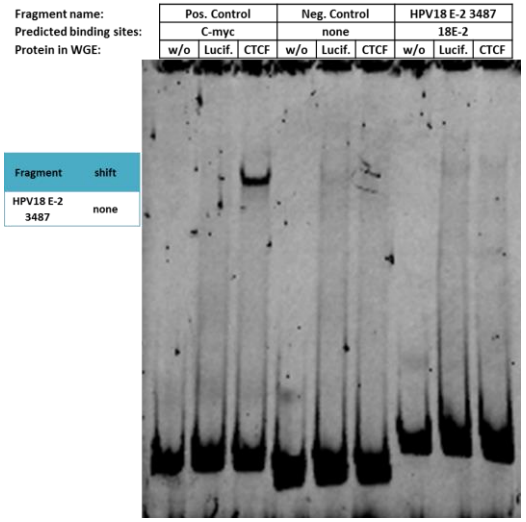
Template	Fragment		Predicted motifs on fragment	Predicted by	Fragment name	CTCF band shift
	Start	End				
HPV18	2926	3117	S 2990, C 2990	1.PWM [18.4], 3.PWM [13.2], 4.PWM [13.3], 4.PWM [7.3], 5.PWM [16.6], 6.PWM [23.4], CTCFBSDB	HPV18 C+1 2990	strong
	1102	1297	1205	Essex	HPV18 E+1 1205	none
	5381	5577	5475	Essex	HPV18 E-1 5475	medium
	3381	3575	3487	Essex	HPV18 E-2 3487	none
	754	943	844	5.PWM [10.0]	HPV18 pred. 844	none
	3527	3718	3621	3.PWM [8.2], 5.PWM [14.0], 6.PWM [14.2]	HPV18 pred. 3621	medium
	4440	4638	S 4505, S 4538	1.PWM [12.2], 5.PWM [10.5]	HPV18 pred. 4505 and 4538	two bands, medium
	5655	5850	5768	1.PWM [12.9], 3.PWM [9.7], 4.PWM [7.3], 5.PWM [10.1], 6.PWM [12.2]	HPV18 pred. 5768	medium
	4947	5155	None	N/A	HPV18 CTCF?A	none
	5045	5253	None	N/A	HPV18 CTCF?B	none

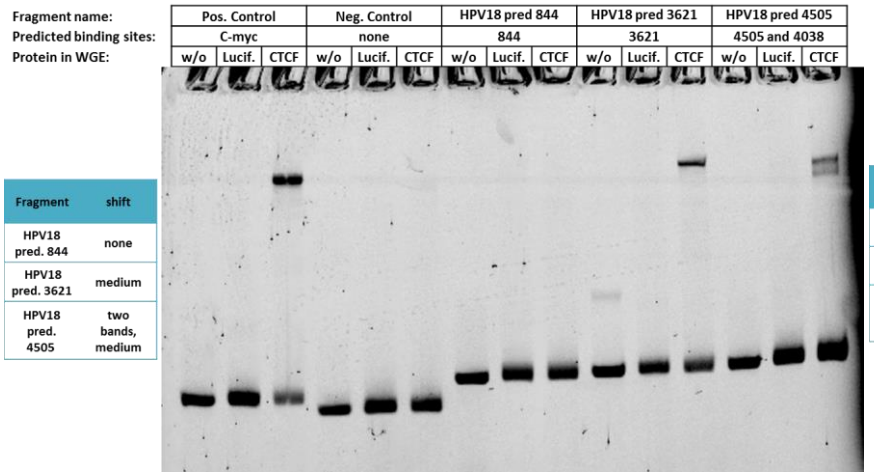
Table 52) HPV18 fragments of the EMSAs images below


Fragment name:
Pos. Control
Neg. Control
HPV18 C+1 2990
HPV18 E+1 1205
HPV18 E-1 5475
Predicted binding sites:
C-myc
none
18C+1
18E+1
18E-1
Protein in WGE:
w/o
Lucif.
CTCF
Fragment
shift
HPV18 C+1 2990
strong
HPV18 E+1 1205
none
HPV18 E-1 5475
medium

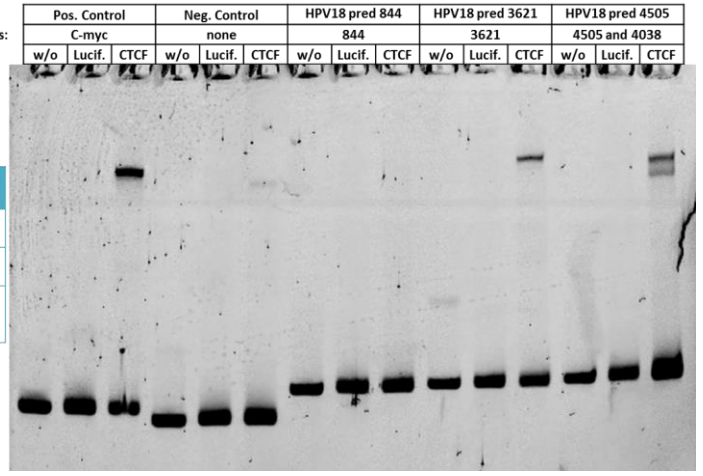

Fragment name:
Pos. Control
Neg. Control
HPV18 C+1 2990
HPV18 E+1 1205
HPV18 E-1 5475
Predicted binding sites:
C-myc
none
18C+1
18E+1
18E-1
Protein in WGE:
w/o
Lucif.
CTCF
Fragment
shift
HPV18 C+1 2990
strong
HPV18 E+1 1205
none
HPV18 E-1 5475
medium


Fragment name:
Pos. Control
Neg. Control
HPV18 E-2 3487
Predicted binding sites:
C-myc
none
18E-2
Protein in WGE:
w/o
Lucif.
CTCF
Fragment
shift
HPV18 E-2 3487
none

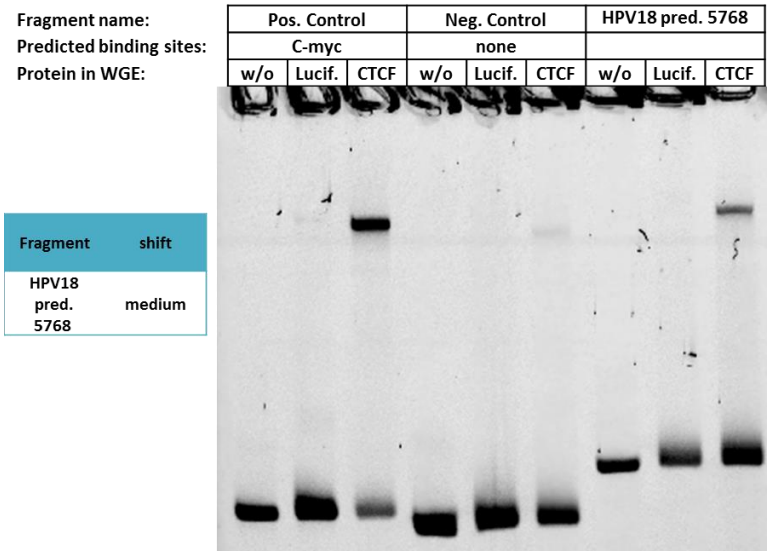

Fragment name:
Pos. Control
Neg. Control
HPV18 E-2 3487
Predicted binding sites:
C-myc
none
18E-2
Protein in WGE:
w/o
Lucif.
CTCF
Fragment
shift
HPV18 E-2 3487
none



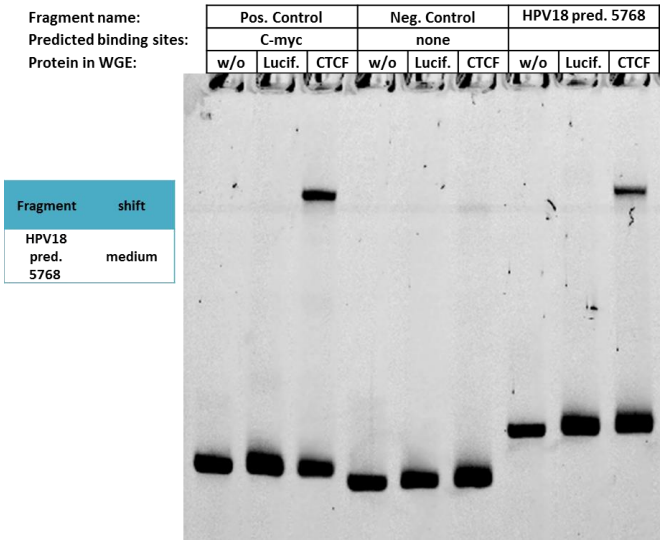
Fragment	shift
HPV18 pred. 844	none
HPV18 pred. 3621	medium
HPV18 pred. 4505	two bands, medium



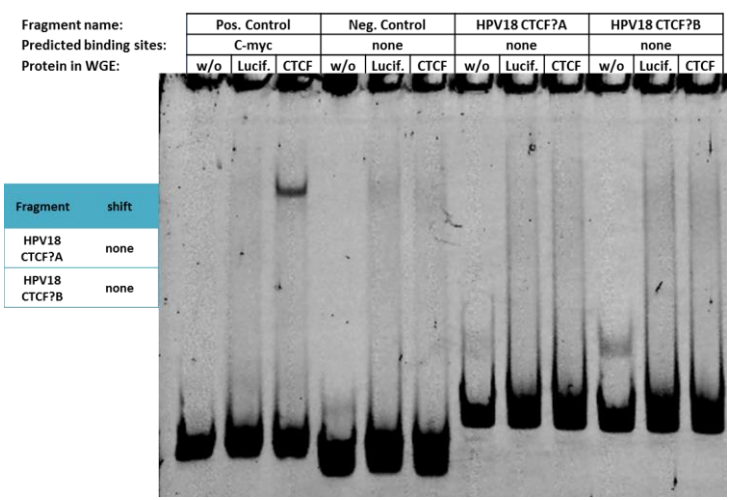
Fragment	shift
HPV18 pred. 844	none
HPV18 pred. 3621	medium
HPV18 pred. 4505	two bands, medium



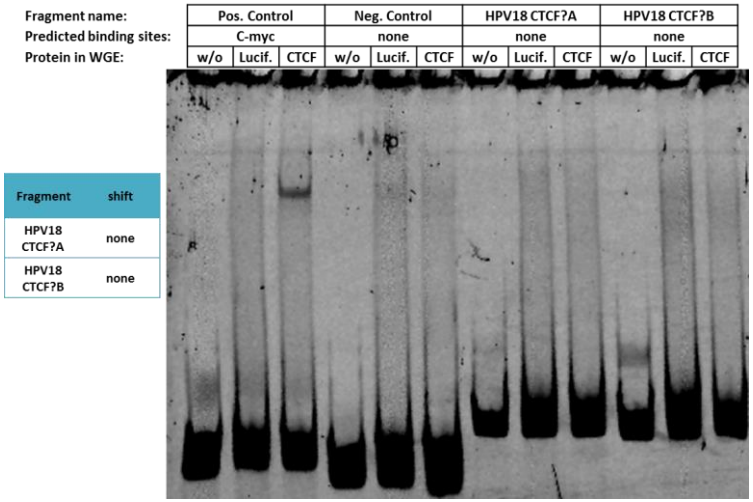
Fragment	shift
HPV18 pred. 5768	medium



Fragment	shift
HPV18 pred. 5768	medium



Fragment	shift
HPV18 CTCF?A	none
HPV18 CTCF?B	none

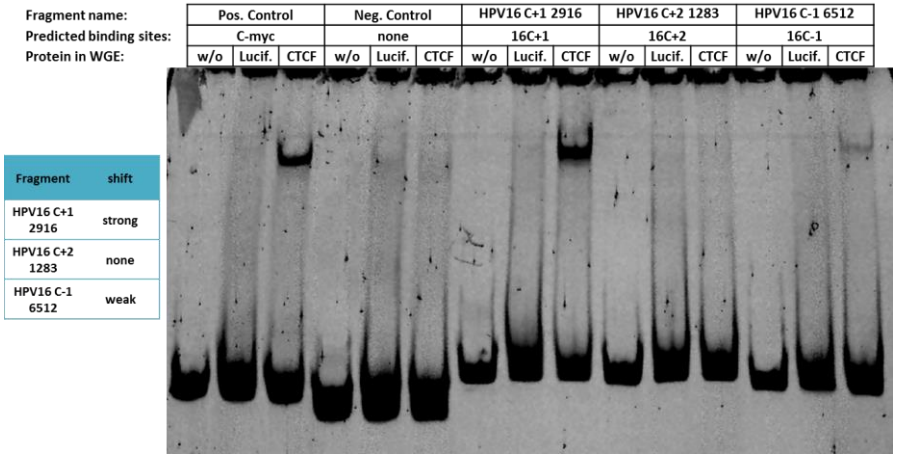


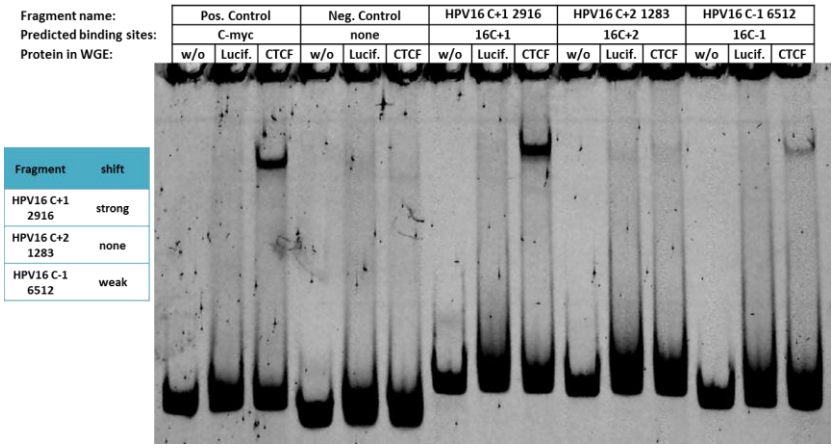
Fragment	shift
HPV18 CTCF?A	none
HPV18 CTCF?B	none

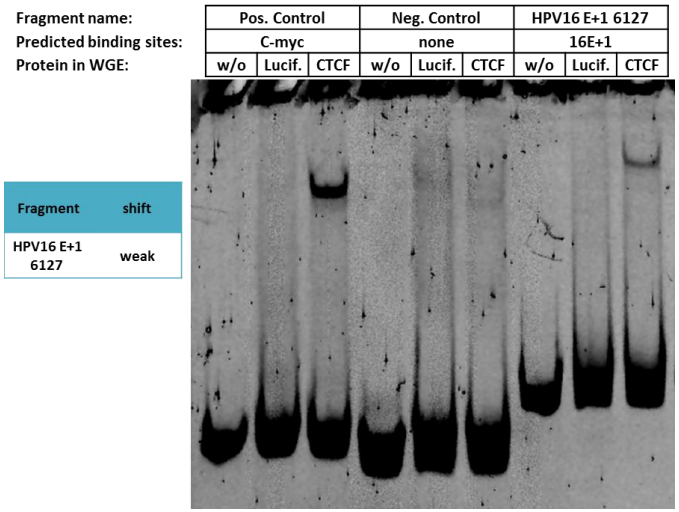
6.4.2.2 HPV16

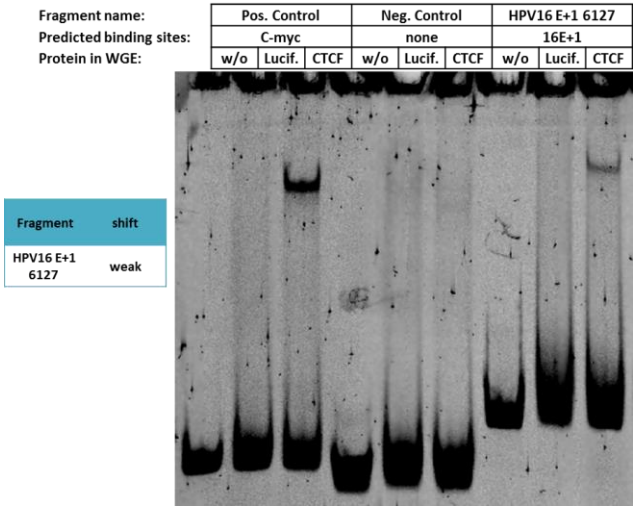
Template	Fragment		Predicted motifs on fragment	Predicted by	Fragment name	CTCF band shift
	Start	End				
HPV16	2852	3049	S 2916, C 2916	3.PWM [12.2], 5.PWM [9.8], 6.PWM [20.1], CTCFBSDB	HPV16 C+1 2916	strong
	1216	1405	S 1283, C 1283	5.PWM [10.2], CTCFBSDB	HPV16 C+2 1283	none
	6426	6600	C 6512, S 6515	1.PWM [13.9], 3.PWM [6.5], 4.PWM [8.6], 6.PWM [9.3], CTCFBSDB	HPV16 C-1 6512	weak
	6051	6278	6127	Essex	HPV16 E+1 6127	weak
	4999	5206	5119	Essex	HPV16 E-1 5119	weak
	6772	6957	6860	1.PWM [12.1], 3.PWM [9.0], 4.PWM [7.7], 6.PWM [9.1]	HPV16 pred. 6860	weak

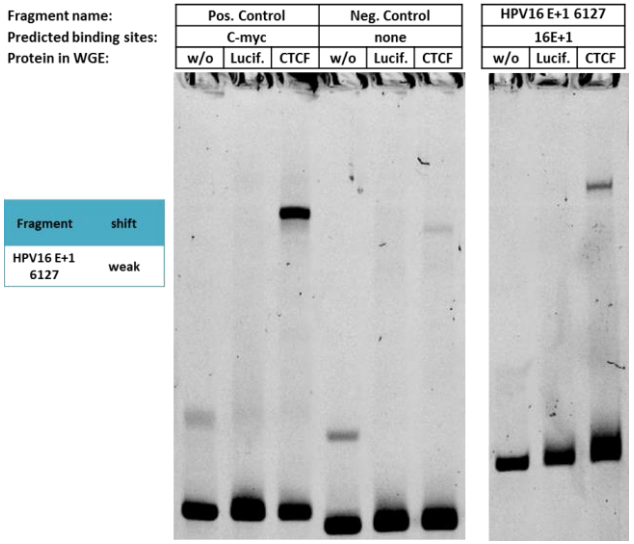
Table 53) HPV16 fragments of the EMSAs images below

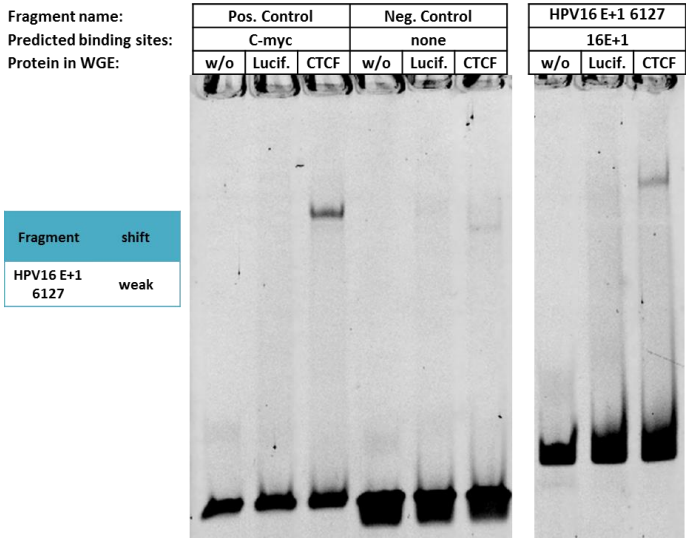


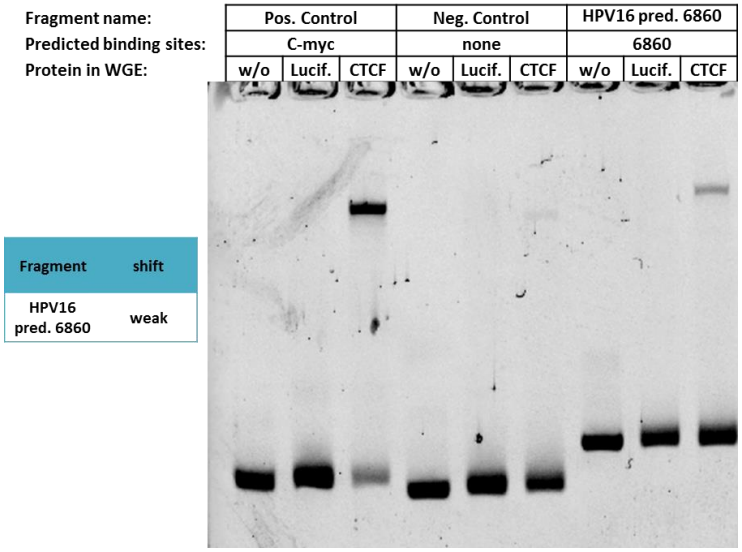



Fragment name:
Predicted binding sites:
Protein in WGE:
Pos. Control
C-myc
w/o Lucif. CTCF
Neg. Control
none
w/o Lucif. CTCF
HPV16 E+1 6127
16E+1
w/o Lucif. CTCF
Fragment shift
HPV16 E+1 6127 weak

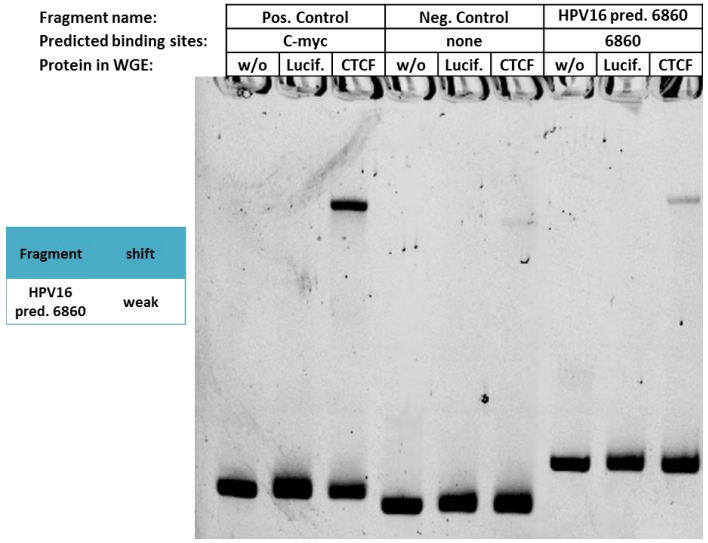

Fragment name:
Predicted binding sites:
Protein in WGE:
Pos. Control
C-myc
w/o Lucif. CTCF
Neg. Control
none
w/o Lucif. CTCF
HPV16 E+1 6127
16E+1
w/o Lucif. CTCF
Fragment shift
HPV16 E+1 6127 weak


Fragment name:
Predicted binding sites:
Protein in WGE:
Pos. Control
C-myc
w/o Lucif. CTCF
Neg. Control
none
w/o Lucif. CTCF
HPV16 E+1 6127
16E+1
w/o Lucif. CTCF
Fragment shift
HPV16 E+1 6127 weak


Fragment name:
Predicted binding sites:
Protein in WGE:
Pos. Control
C-myc
w/o Lucif. CTCF
Neg. Control
none
w/o Lucif. CTCF
HPV16 E+1 6127
16E+1
w/o Lucif. CTCF
Fragment shift
HPV16 E+1 6127 weak



Fragment	shift
HPV16 pred. 6860	weak

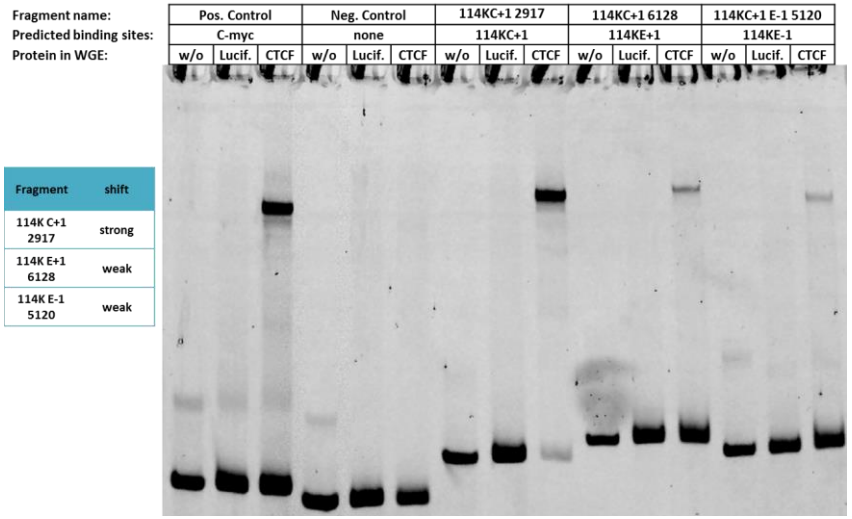


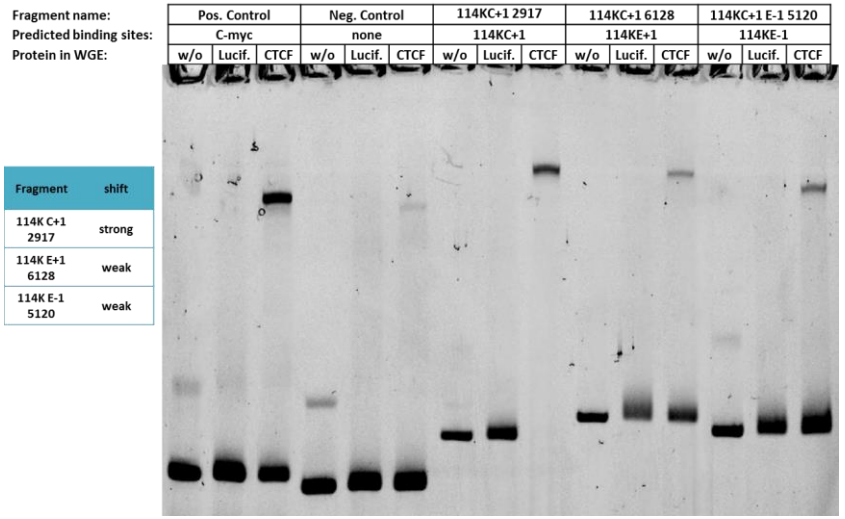
Fragment	shift
HPV16 pred. 6860	weak

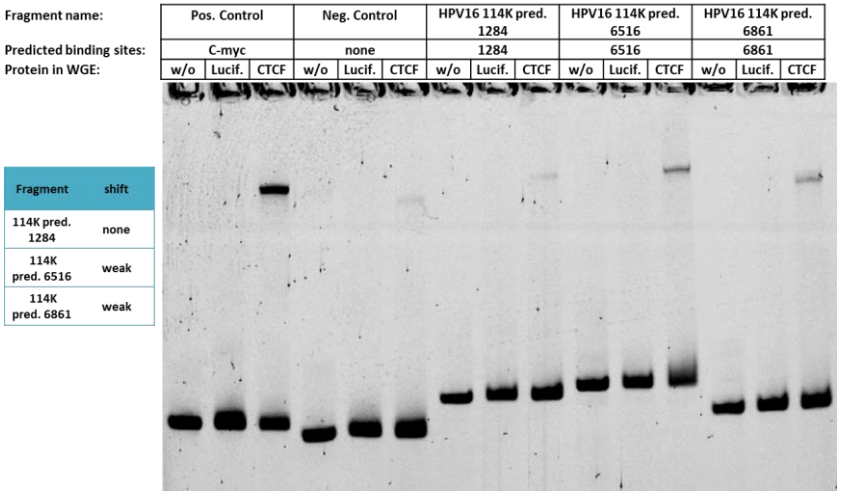
6.4.2.3 HPV16 114/K

Template	Fragment Start	Fragment End	Predicted motifs on fragment	Predicted by	Fragment name	CTCF band shift
HPV16 114/K	2852	3049	S 2917, C 2917	1.PWM [19.3], 3.PWM [22.4], 4.PWM [16.0], 5.PWM [18.2], 6.PWM [27.6], CTCFBSDB	114K C+1 2917	strong
	6051	6278	6128	Essex	114K E+1 6128	weak
	5000	5207	5120	Essex	114K E-1 5120	weak
	1216	1405	1284	5.PWM [10.2]	114K pred. 1284	none
	6421	6621	6516	1.PWM [13.9], 3.PWM [6.5], 4.PWM [8.6]	114K pred. 6516	weak
	6783	6945	6861	1.PWM [12.1], 3.PWM [8.9], 4.PWM [7.7], 6.PWM [9.1]	114K pred. 6861	weak

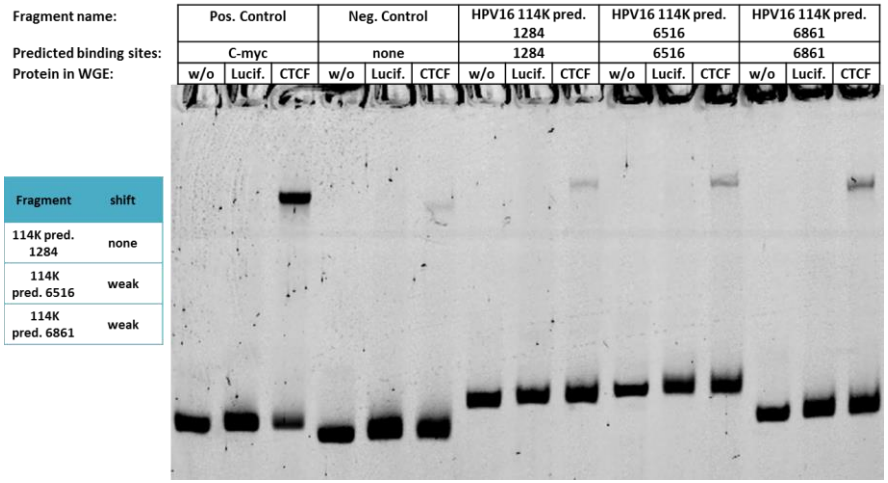
Table 54) HPV16 114/K fragments of the EMSAs images below







Fragment	shift
114K pred. 1284	none
114K pred. 6516	weak
114K pred. 6861	weak



Fragment	shift
114K pred. 1284	none
114K pred. 6516	weak
114K pred. 6861	weak

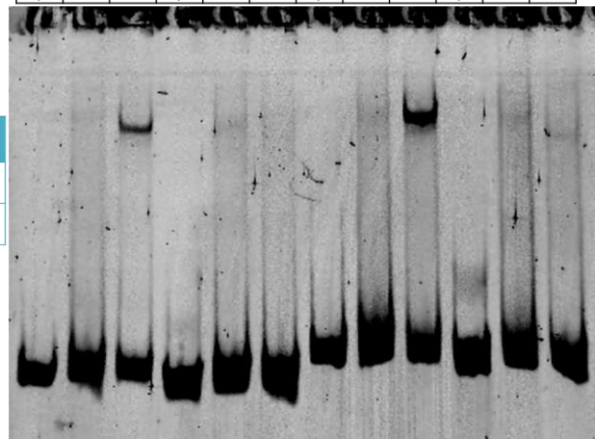
6.4.2.4 HPV31

Template	Fragment		Predicted motifs on fragment	Predicted by	Fragment name	CTCF band shift
	Start	End				
HPV31	2357	2531	S 2413, C 2413	1.PWM [14.4], 4.PWM [14.7], CTCFBSDB	HPV31 C+1 2413	strong
	1182	1374	C 1278 , S 1278, E 1294	CTCFBSDB, Essex, 5.PWM [14.0]	HPV31 C+2/E+3 1278	none
	2230	2406	2332	CTCFBSDB	HPV31 C+3 2332	none
	5077	5273	5179	Essex	HPV31 E+1 5179	strong
	1029	1200	1093	Essex	HPV31 E+2 1093	none
	804	1008	885	Essex	HPV31 E+4 885	none
	534	713	616	4.PWM [8.2]	HPV31 pred. 616	none
	6354	6540	6432	1.PWM [13.5], 3.PWM [7.2]	HPV31 pred. 6432	medium
	2801	3015	2854	1.PWM [19.2], 3.PWM [22.3], 4.PWM [16.0], 4.PWM [9.0], 5.PWM [18.9], 6.PWM [27.5]	HPV31 CTCF?A	strong
	2894	3093	None	N/A	HPV31 CTCF?B	none

Table 55) HPV31 fragments of the EMSAs images below

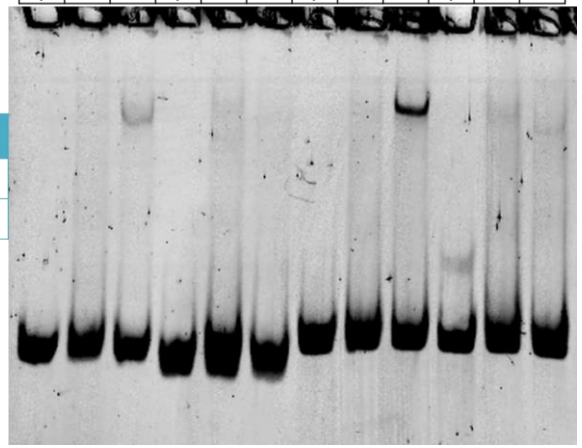
Fragment name:	Pos. Control			Neg. Control			HPV31 C+1 2413			HPV31 C+3 2332		
Predicted binding sites:	C-myc			none			31C+1			31C+3		
Protein in WGE:	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF

Fragment	shift
HPV31 C+1 2413	strong
HPV31 C+3 2332	none



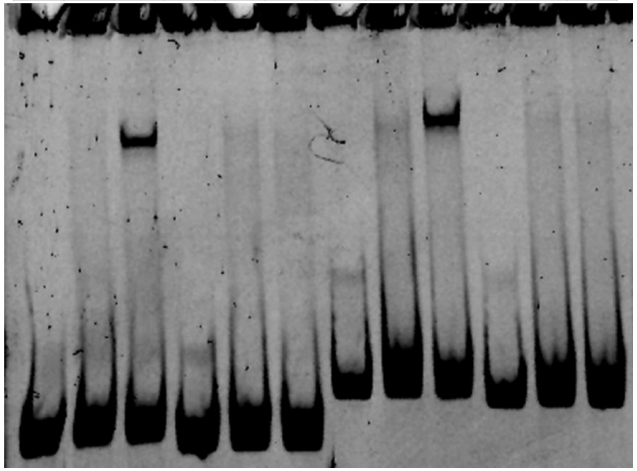
Fragment name:	Pos. Control			Neg. Control			HPV31 C+1 2413			HPV31 C+3 2332		
Predicted binding sites:	C-myc			none			31C+1			31C+3		
Protein in WGE:	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF

Fragment	shift
HPV31 C+1 2413	strong
HPV31 C+3 2332	none



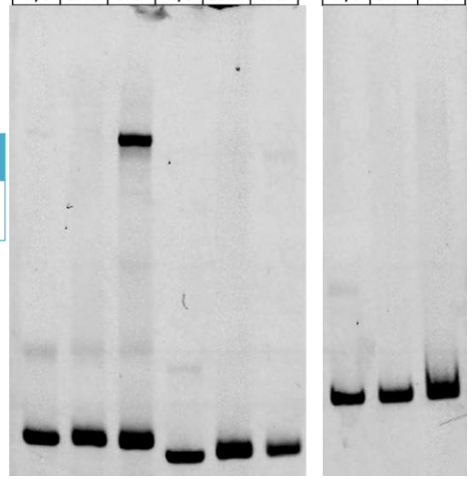
Fragment name:	Pos. Control			Neg. Control			HPV31 C+1 2413			HPV31 C+2/E+3 1278		
Predicted binding sites:	C-myc			none			31C+1			31C+2/31E+3		
Protein in WGE:	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF

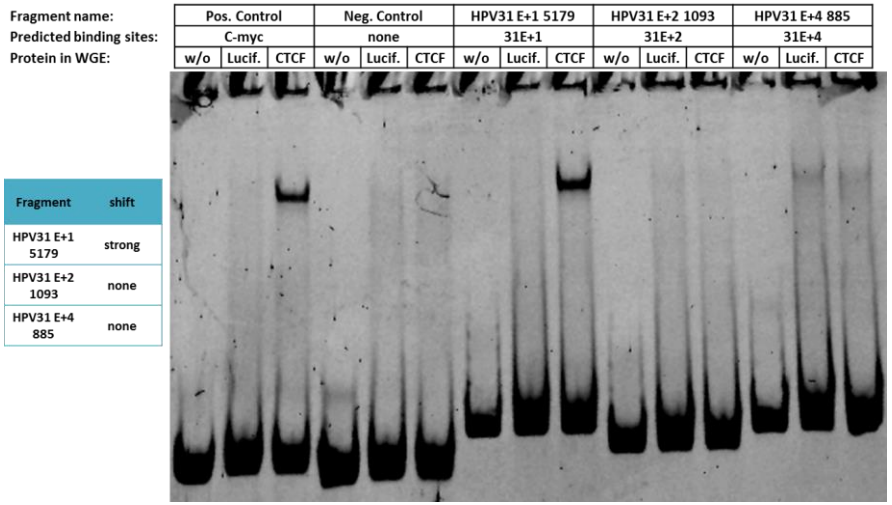
Fragment	shift
HPV31 C+1 2413	strong
HPV31 C+2/E+3 1278	none



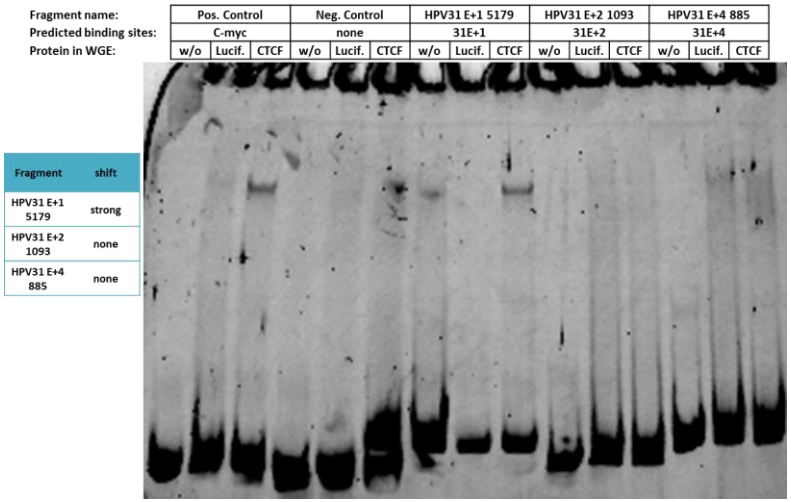
Fragment name:	Pos. Control			Neg. Control			HPV31 C+2/E+3 1278		
Predicted binding sites:	C-myc			none			31C+2/31E+3		
Protein in WGE:	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF

Fragment	shift
HPV31 C+2/E+3 1278	none

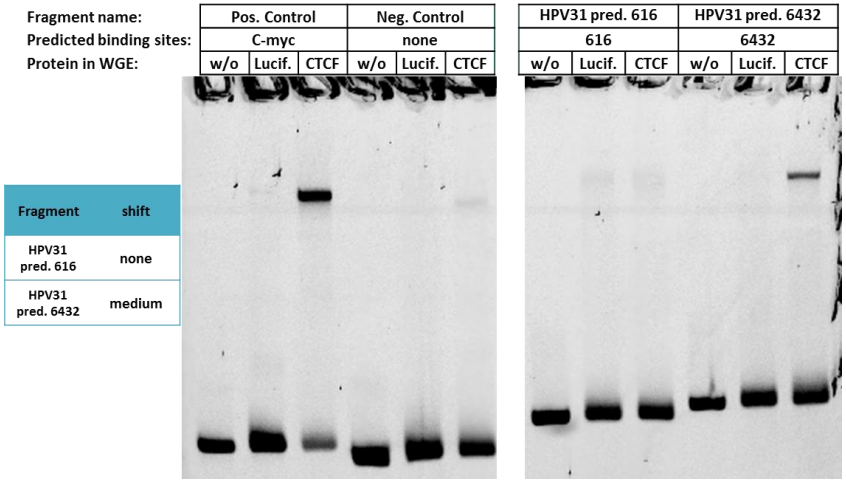




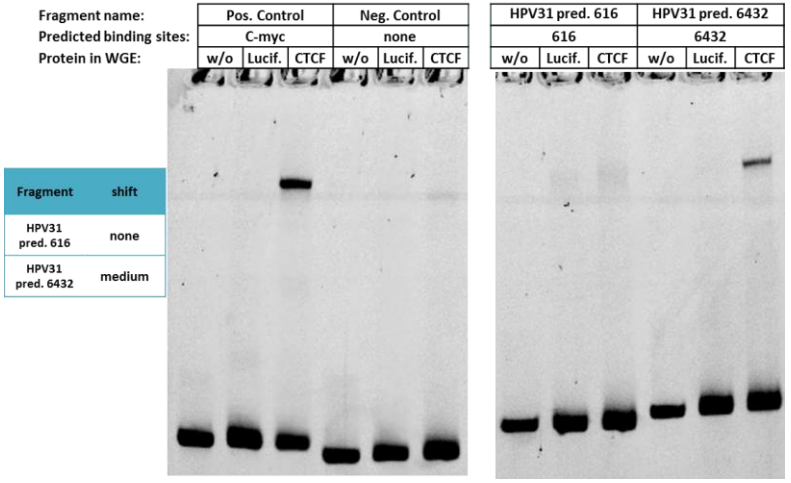
Fragment	shift
HPV31 E+1 5179	strong
HPV31 E+2 1093	none
HPV31 E+4 885	none



Fragment	shift
HPV31 E+1 5179	strong
HPV31 E+2 1093	none
HPV31 E+4 885	none

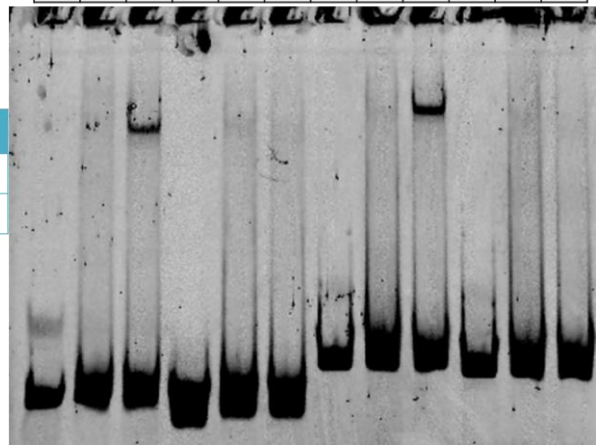


Fragment	shift
HPV31 pred. 616	none
HPV31 pred. 6432	medium



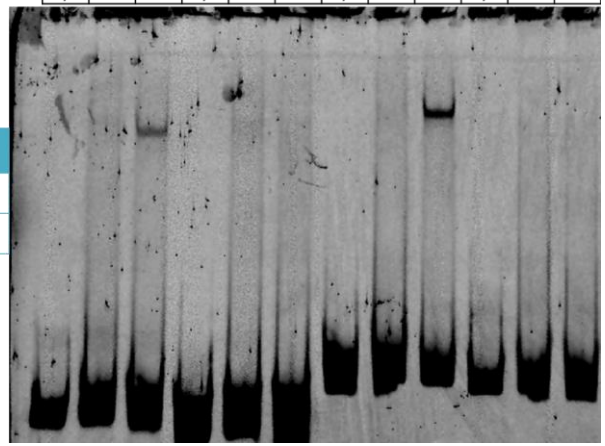
Fragment	shift
HPV31 pred. 616	none
HPV31 pred. 6432	medium

Fragment name:	Pos. Control			Neg. Control			HPV31 CTCF?A			HPV31 CTCF?B		
Predicted binding sites:	C-myc			none			2854			none		
Protein in WGE:	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF



Fragment	shift
HPV31 CTCF?A	strong
HPV31 CTCF?B	none

Fragment name:	Pos. Control			Neg. Control			HPV31 CTCF?A			HPV31 CTCF?B		
Predicted binding sites:	C-myc			none			2854			none		
Protein in WGE:	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF



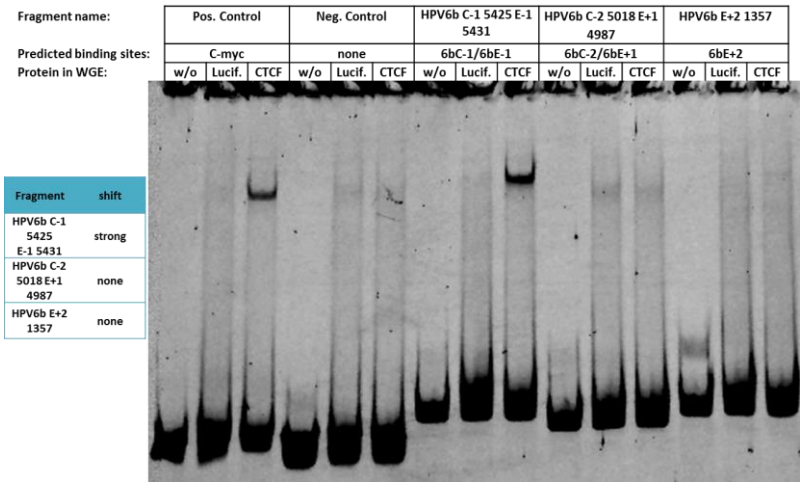
Fragment	shift
HPV31 CTCF?A	strong
HPV31 CTCF?B	none

6.4.3 Low risk HPV types

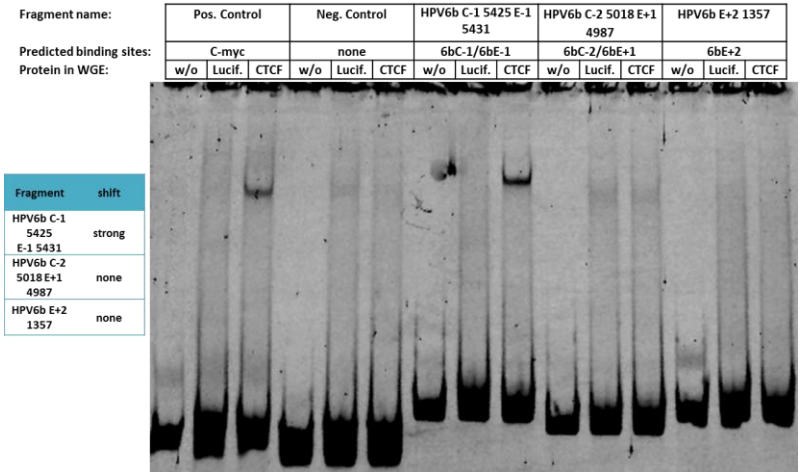
6.4.3.1 HPV6b

Template	Fragment		Predicted motifs on fragment	Predicted by	Fragment name	CTCF band shift
	Start	End				
HPV6b	5317	5515	S 5425, C 5425, E 5431	1.PWM [16.9], 3.PWM [17.0], 4.PWM [6.9], 5.PWM [12.9], 6.PWM [12.2], CTCFBSDB, Essex	HPV6b C-1 5425 E-1 5431	strong
	4913	5102	S 5018, C 5018, E 4987	4.PWM [7.5], CTCFBSDB, Essex	HPV6b C-2 5018 E+1 4987	none
	1251	1460	1357	Essex	HPV6b E+2 1357	none
	5995	6199	S 6110, E 6099	5.PWM [10.0], Essex	HPV6b E+3 6099	none
	6179	6382	6264	Essex	HPV6b E-2 6264	medium
	7155	7380	E 7206, E 7257	Essex	HPV6b E-3 7206 E-4 7257	none
	4715	4913	4790	Essex	HPV6b E-5 4790	none
	2801	3007	none	N/A	HPV6b CTCF?A	none
	2887	3101	none	N/A	HPV6b CTCF?B	none

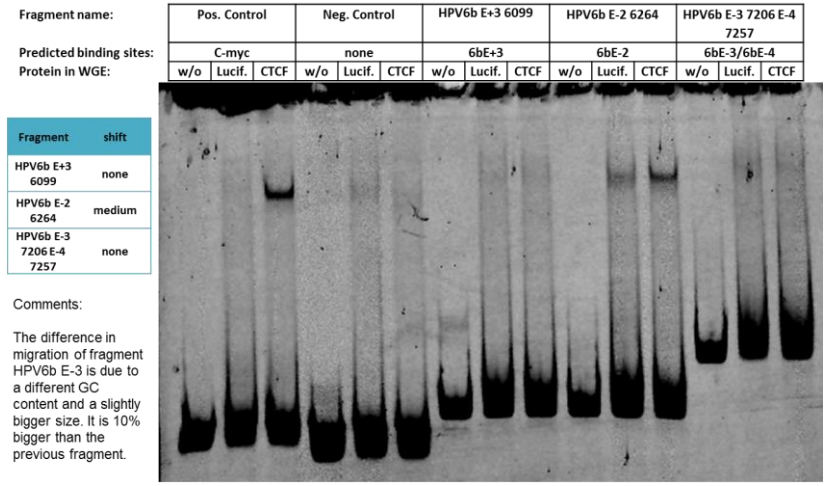
Table 56) HPV6b fragments of the EMSAs images below



Fragment	shift
HPV6b C-1 5425 E-1 5431	strong
HPV6b C-2 5018 E+1 4987	none
HPV6b E+2 1357	none



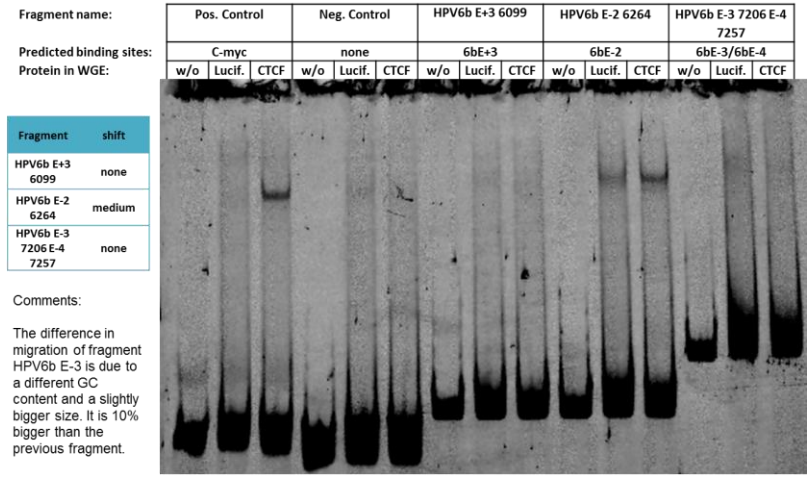
Fragment	shift
HPV6b C-1 5425 E-1 5431	strong
HPV6b C-2 5018 E+1 4987	none
HPV6b E+2 1357	none



Fragment	shift
HPV6b E+3 6099	none
HPV6b E-2 6264	medium
HPV6b E-3 7206 E-4 7257	none

Comments:

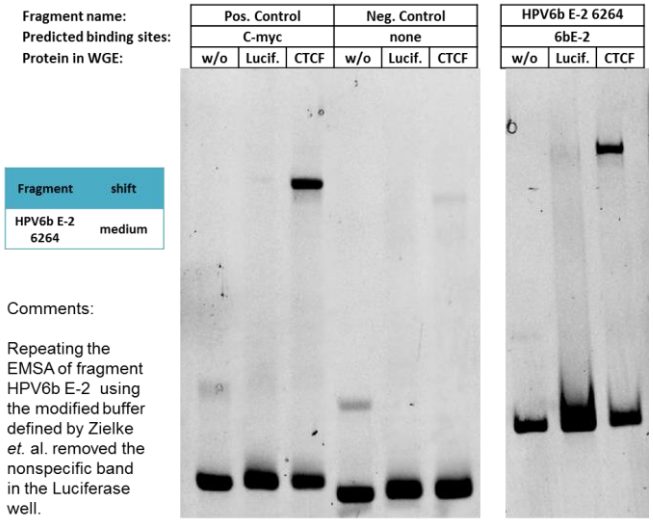
The difference in migration of fragment HPV6b E-3 is due to a different GC content and a slightly bigger size. It is 10% bigger than the previous fragment.



Fragment	shift
HPV6b E+3 6099	none
HPV6b E-2 6264	medium
HPV6b E-3 7206 E-4 7257	none

Comments:

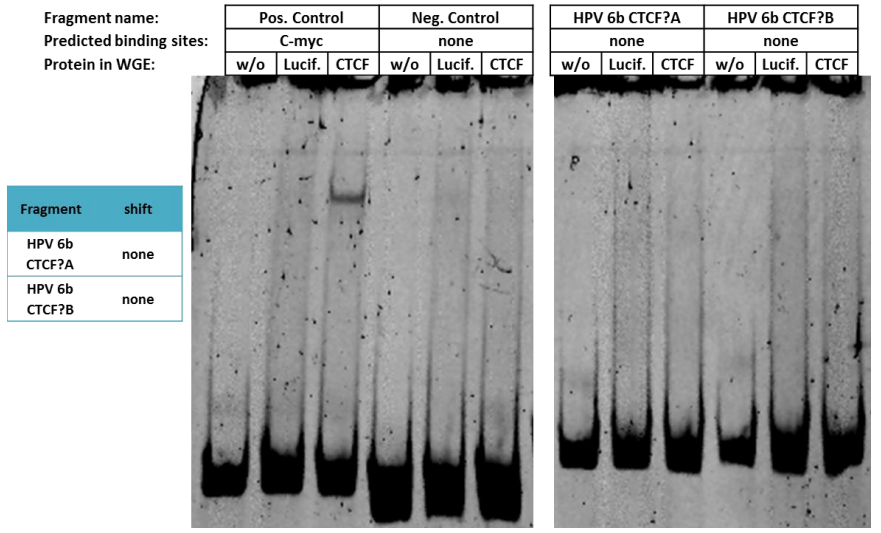
The difference in migration of fragment HPV6b E-3 is due to a different GC content and a slightly bigger size. It is 10% bigger than the previous fragment.



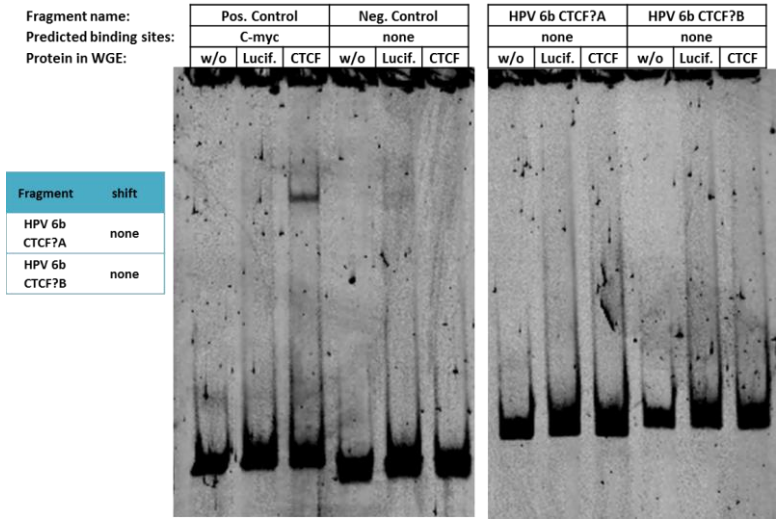
Fragment	shift
HPV6b E-2 6264	medium

Comments:

Repeating the EMSA of fragment HPV6b E-2 using the modified buffer defined by Zielke *et.* al. removed the nonspecific band in the Luciferase well.



Fragment	shift
HPV 6b CTCF?A	none
HPV 6b CTCF?B	none

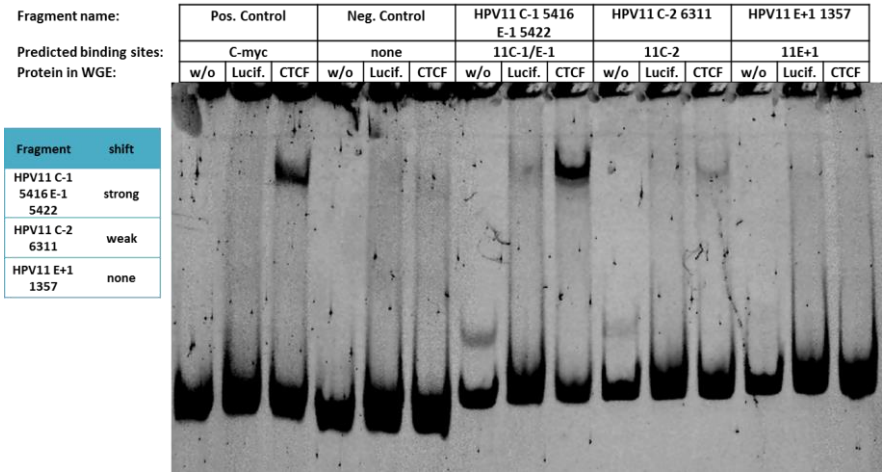


Fragment	shift
HPV 6b CTCF?A	none
HPV 6b CTCF?B	none

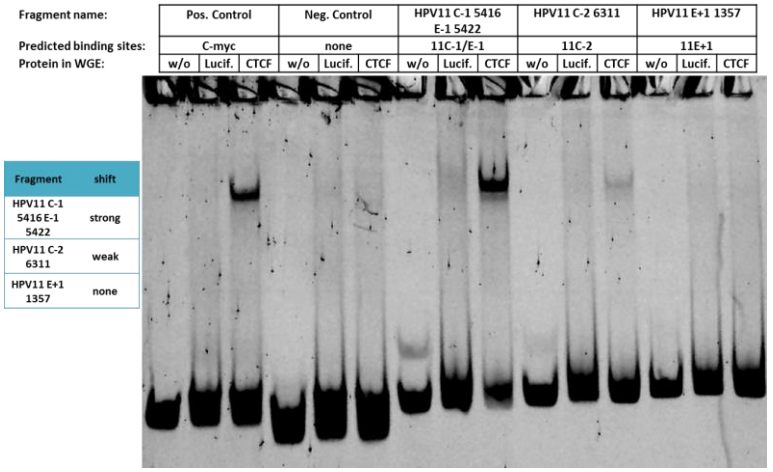
6.4.3.2 HPV11

Template	Fragment Start	Fragment End	Predicted motifs on fragment	Predicted by	Fragment name	CTCF band shift
HPV11	5330	5501	S 5416, C 5416, E 5422	1.PWM [18.4], 3.PWM [20.5], 4.PWM [11.5], 5.PWM [18.6], 6.PWM [15.8], CTCFBSDB, Essex	HPV11 C-1 5416 E-1 5422	strong
	6243	6428	S 6311, C 6311	4.PWM [13.2], CTCFBSDB	HPV11 C-2 6311	weak
	1295	1494	1357	Essex	HPV11 E+1 1357	none
	4709	4898	4781	Essex	HPV11 E-2 4781	none
	3930	4153	4058	Essex	HPV11 E-3 4058	medium
	4844	5041	4921	4.PWM [7.3]	HPV11 pred. 4921	weak
	6544	6738	6636	4.PWM [7.3]	HPV11 pred. 6636	weak, smear
	6872	7074	6980	5.PWM [10.1], 6.PWM [9.3]	HPV11 pred. 6980	none
	2801	3003	None	N/A	HPV11 CTCF?A	none
	2900	3104	None	N/A	HPV11 CTCF?B	none

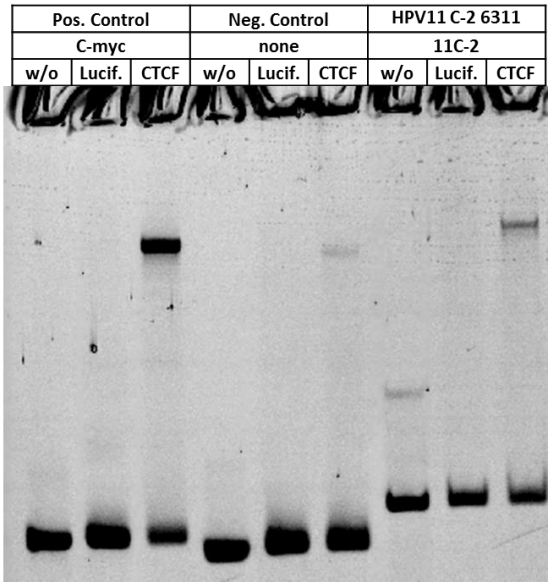
Table 57) HPV11 fragments of the EMSAs images below



Fragment	shift
HPV11 C-1 5416 E-1 5422	strong
HPV11 C-2 6311	weak
HPV11 E+1 1357	none



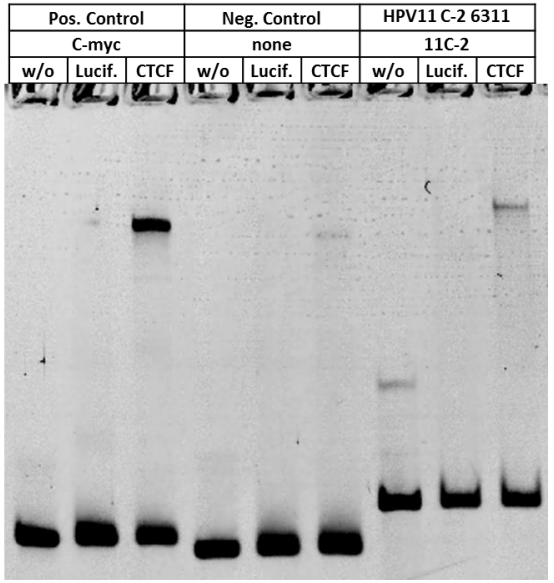
Fragment	shift
HPV11 C-1 5416 E-1 5422	strong
HPV11 C-2 6311	weak
HPV11 E+1 1357	none



Fragment	shift
HPV11 C-2 6311	weak

Comments:

Repeating the EMSA of fragment HPV11 C-2 using the modified buffer defined by Zielke *et.* al. resulted in a more clear band.

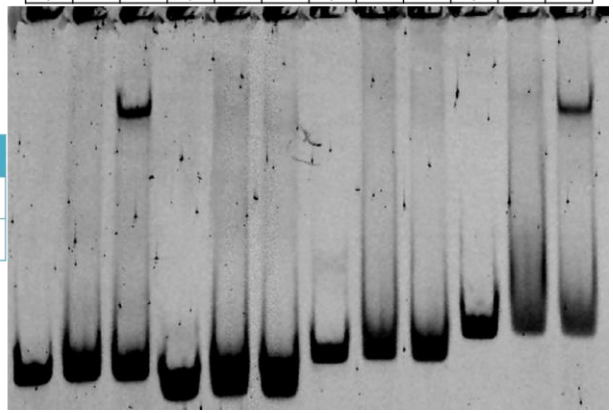


Fragment	shift
HPV11 C-2 6311	weak

Comments:

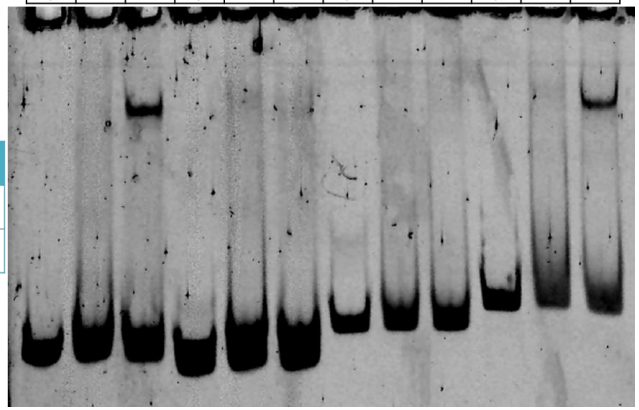
Repeating the EMSA of fragment HPV11 C-2 using the modified buffer defined by Zielke *et.* al. resulted in a more clear band.

Fragment name:	Pos. Control			Neg. Control			HPV11 E-2 4781			HPV11 E-3 4058		
Predicted binding sites:	C-myc			none			11E-2			11E-3		
Protein in WGE:	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF



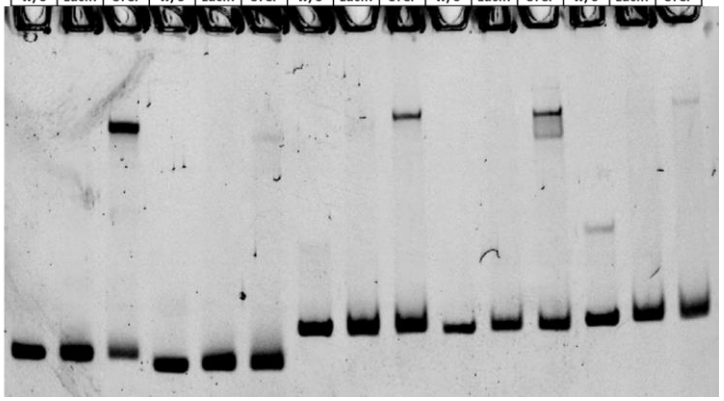
Fragment	shift
HPV11 E-2 4781	none
HPV11 E-3 4058	medium

Fragment name:	Pos. Control			Neg. Control			HPV11 E-2 4781			HPV11 E-3 4058		
Predicted binding sites:	C-myc			none			11E-2			11E-3		
Protein in WGE:	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF



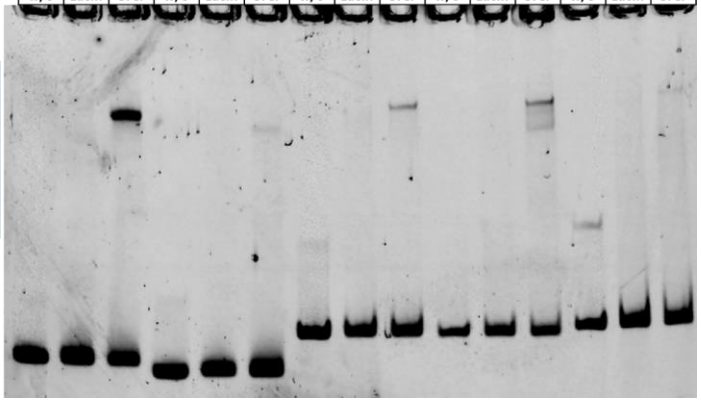
Fragment	shift
HPV11 E-2 4781	none
HPV11 E-3 4058	medium

Fragment name:	Pos. Control			Neg. Control			HPV11 pred. 4921			HPV11 pred. 6636			HPV11 pred. 6980		
Predicted binding sites:	C-myc			none			4921			6636			6980		
Protein in WGE:	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF

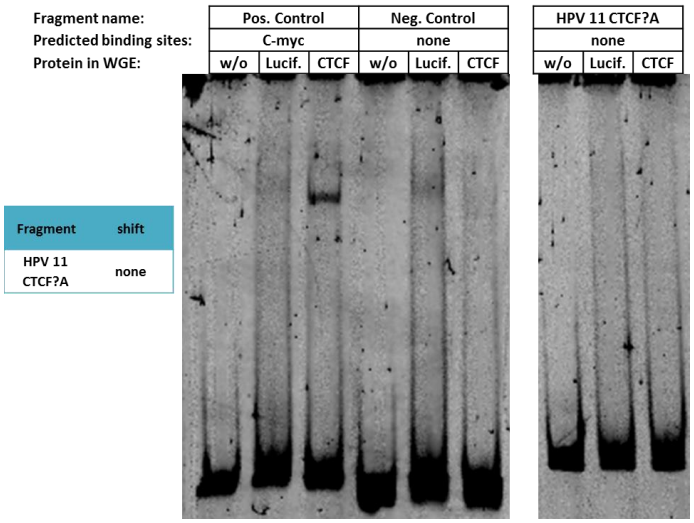


Fragment	shift
HPV11 pred. 4921	weak
HPV11 pred. 6636	two bands weak
HPV11 pred. 6980	none

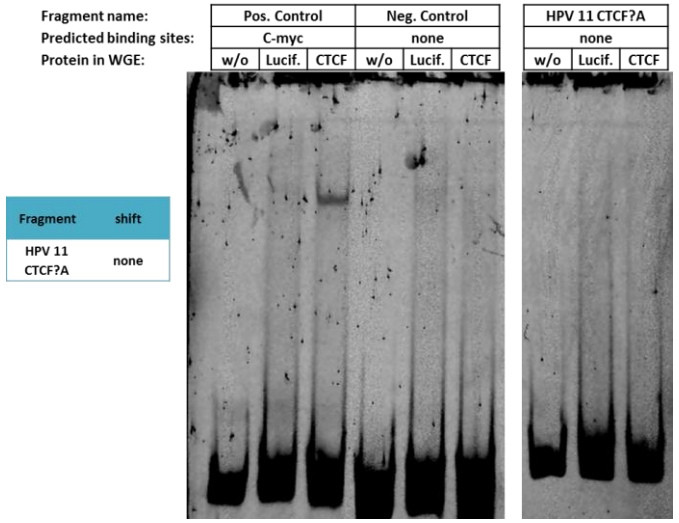
Fragment name:	Pos. Control			Neg. Control			HPV11 pred. 4921			HPV11 pred. 6636			HPV11 pred. 6980		
Predicted binding sites:	C-myc			none			4921			6636			6980		
Protein in WGE:	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF	w/o	Lucif.	CTCF



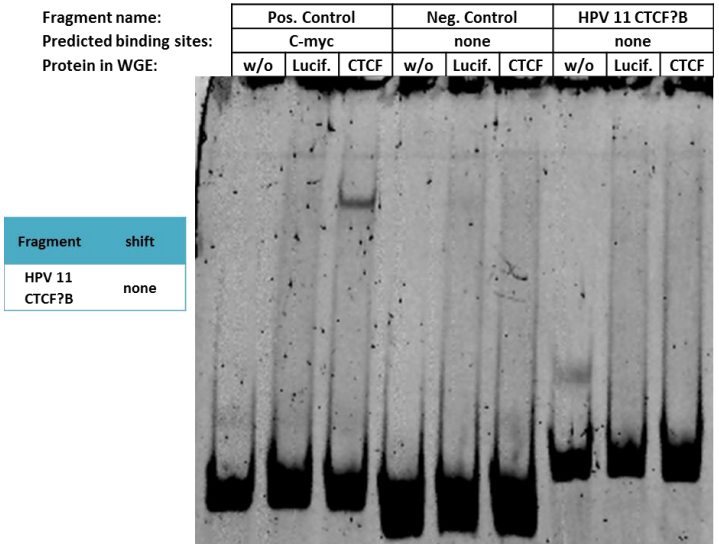
Fragment	shift
HPV11 pred. 4921	weak
HPV11 pred. 6636	two bands weak
HPV11 pred. 6980	none



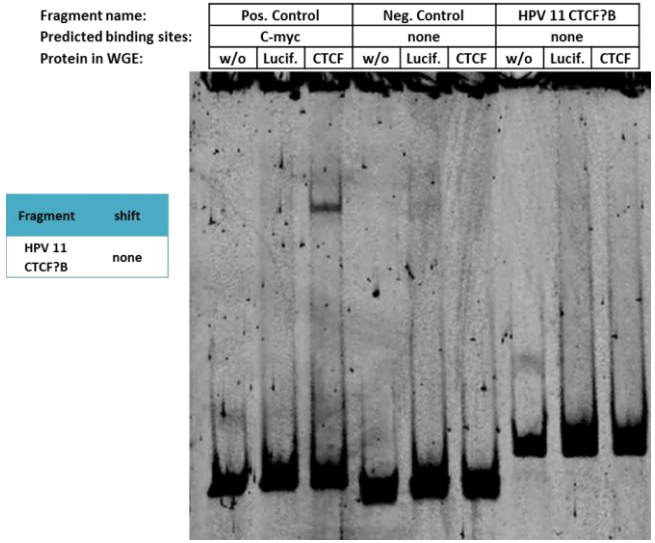
Fragment	shift
HPV 11 CTCF?A	none



Fragment	shift
HPV 11 CTCF?A	none



Fragment	shift
HPV 11 CTCF?B	none

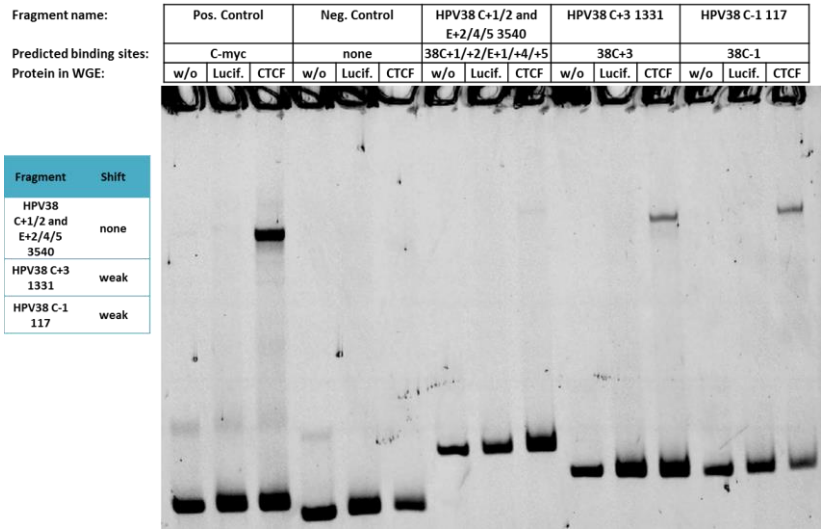


Fragment	shift
HPV 11 CTCF?B	none

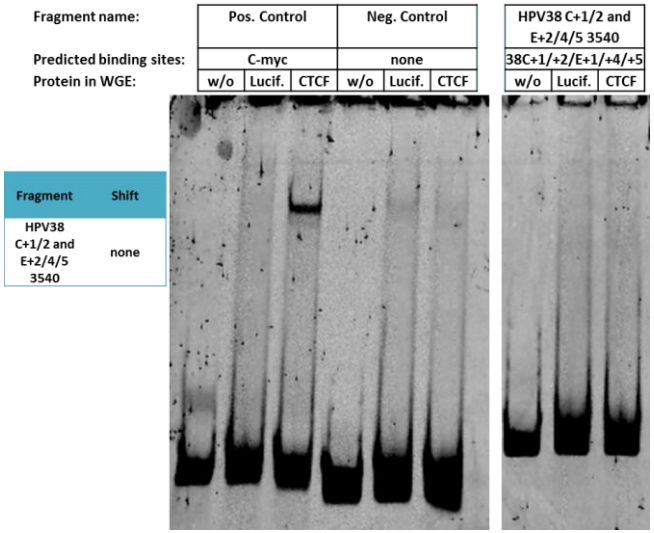
6.4.4 HPV38

Template	Fragment Start	Fragment End	Predicted motifs on fragment	Predicted by	Fragment name	CTCF band shift
HPV38	3436	3666	S 3543, S 3573, C 3540, C 3571, E 3520, E 3544 E 3487	1.PWM [13.3], 1.PWM [15.3], 5.PWM [9.2], CTCFBSDB, Essex	HPV38 C+1/2 and E+2/4/5 3540	none
	1243	1431	S 1331, C 1331	1.PWM [14.9], 5.PWM [15.6], 6.PWM [8.5], CTCFBSDB	HPV38 C+3 1331	medium
	35	219	S 117, C 117	6.PWM [10.9], CTCFBSDB	HPV38 C-1 117	medium
	176	363	254	Essex	HPV38 E+1 254	none
	4357	4545	4454	Essex	HPV38 E+3 4454	none
	584	791	690	6PWM [9.2]	HPV38 pred. 690	none

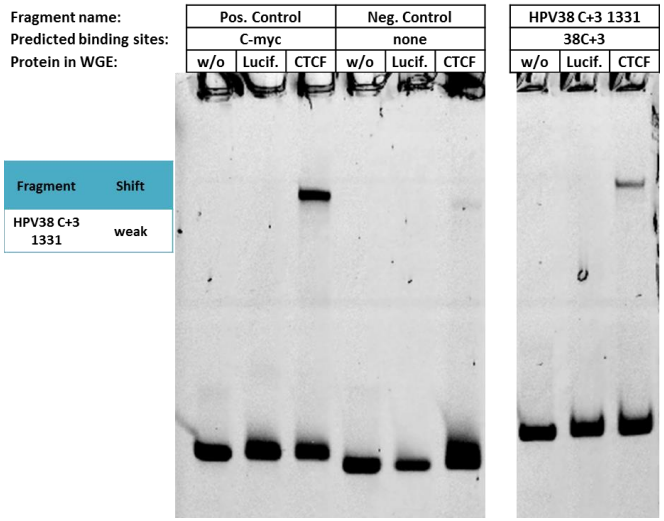
Table 58) HPV38 fragments of the EMSAs images below



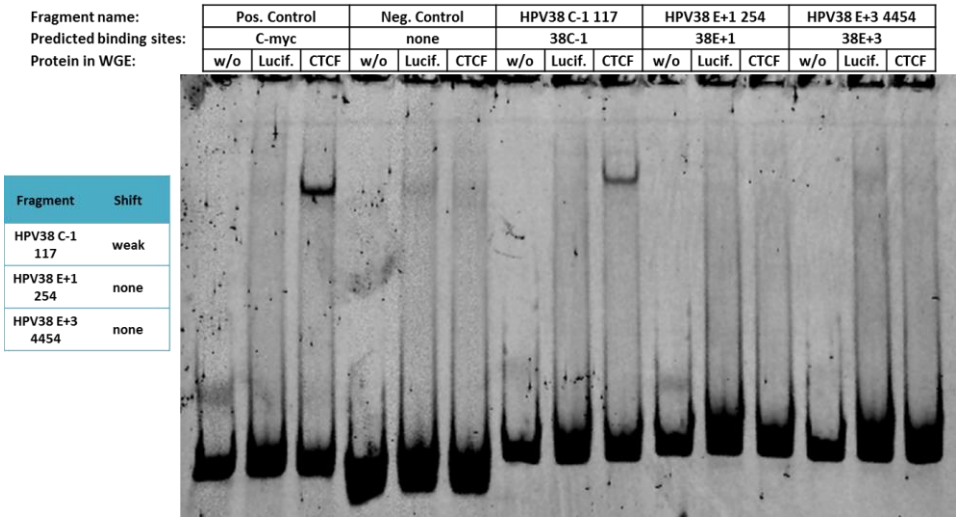
Fragment	Shift
HPV38 C+1/2 and E+2/4/5 3540	none
HPV38 C+3 1331	weak
HPV38 C-1 117	weak



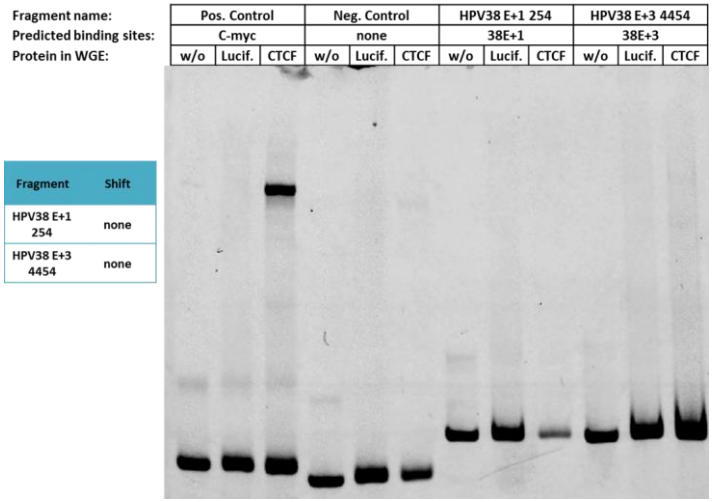
Fragment	Shift
HPV38 C+1/2 and E+2/4/5 3540	none



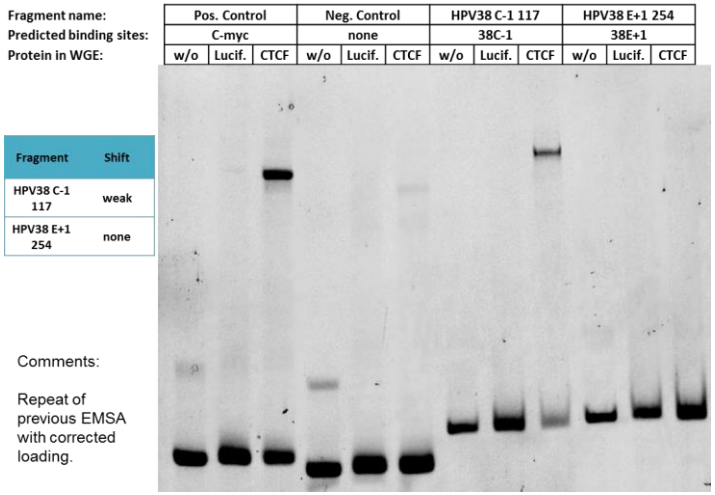
Fragment	Shift
HPV38 C+3 1331	weak



Fragment	Shift
HPV38 C-1 117	weak
HPV38 E+1 254	none
HPV38 E+3 4454	none



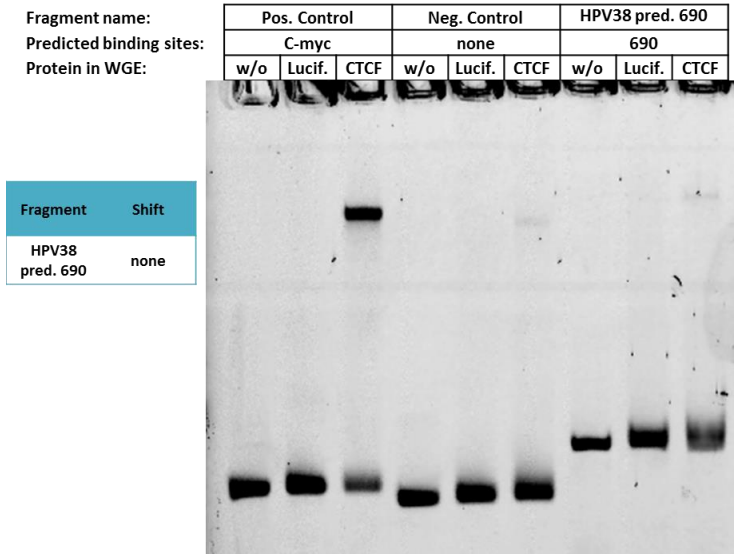
Fragment	Shift
HPV38 E+1 254	none
HPV38 E+3 4454	none



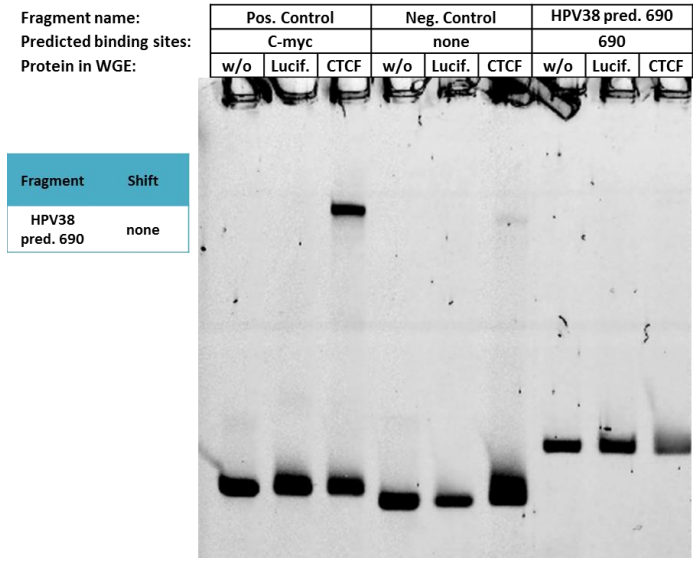
Fragment	Shift
HPV38 C-1 117	weak
HPV38 E+1 254	none

Comments:

Repeat of previous EMSA with corrected loading.



Fragment	Shift
HPV38 pred. 690	none



Fragment	Shift
HPV38 pred. 690	none